



Statistical Learning of **Distributionally Robust Stochastic Control** in Continuous State Spaces

Shengbo Wang (Stanford → USC)
Joint work with *Jason Meng, Nian Si,*
Jose Blanchet, and Zhengyuan Zhou

This work is supported generously by NSF 2229012, 2312204, 2312205, 2403007,
and Air Force Office of Scientific Research FA9550-20-1-0397, and ONR 1398311.



Motivation

Controlled Stochastic Dynamics

$$X_{t+1} = f(X_t, A_t, W_t)$$

State transition function: f

State: $x, X_t \in \mathbb{X} \subset \mathbb{R}^{d_{\mathbb{X}}}$

Action: $A_t \in \mathbb{A}$

Noise: $W_t \in \mathbb{W} \subset \mathbb{R}^{d_{\mathbb{W}}}$

$$E_x^\pi \sum_{t=0}^{\infty} \alpha^t r(X_t, A_t, W_t)$$

Control policy: $\pi \in \Pi$

Reward function: r

$\{W_t : t \geq 0\}$ are i.i.d.

$$X_{t+1} = (X_t + A_t - W_t)_+$$

A simple inventory model



Systems in Operations Research

$$X_{t+1} = f(X_t, A_t, W_t)$$

f is known



Data for W_t is available

Systems in Operations Research

$$W_t \stackrel{d}{=} D \text{ i.i.d.}?$$

In dynamic decision-making context,
model misspecification due to:

- Distribution shifts.
- Temporal correlation within $\{W_t : t \geq 0\}$.
E.g. AR(1) $W_t = D_t + \delta W_{t-1}$
- Input might depend on history.
E.g. $W_t = D_t + \delta g(X_t, A_t, X_{t-1} \dots)$

where $\{D_t : t \geq 0\}$ an i.i.d. sequence.



$$X_{t+1} = (X_t + A_t - W_t)_+$$

Systems in Operations Research

$$W_t \stackrel{d}{=} D \text{ i.i.d.}$$

In dynamic decision
model misspecification

*Learn good and reliable dynamic decisions
that are robust to these risks,
with statistical guarantees.*

E.g.

- Input mismatch

E.g. $W_t = D_t + \epsilon_t$

where $\{D_t : t \geq 0\}$ an i.i.d. sequence

$$X_{t+1} = (X_t + A_t - W_t)_+$$

Literature on DR Policy Learning

Sample complexity of DRRL in finite state and action spaces: [Zhou et al. 21], [Panaganti and Kalathil 21], [Yang et al. 22], [[W](#) et al. 23a], [Shi et al. 23], [[W](#) et al. 23b], [Shi and Chi 24].....

DR Contextual Bandit: [Mu et al. 22], [Si et al. 23], [Shen et al. 23]

Linear and or kernel based DRRL in continuous state spaces: [Blanchet et al. 23], [Ma et al. 22]

Statistical analysis of DR single stage optimal decisions: [Duchi and Namkoong 21], [Lee and Raginsky 18]

Statistical analysis of DR stochastic control in continuous state spaces: [[W](#) et al. 24b] (this paper)



Formulation

Adversarial Robustness Approach

$$\sup_{\pi \in \Pi} E_x^\pi \left[\sum_{t=0}^{\infty} \alpha^t r(X_t, A_t, W_t) \right] \text{ vs. } \sup_{\pi \in \Pi} \inf_{\gamma \in \Gamma} E_x^{\pi, \gamma} \left[\sum_{t=0}^{\infty} \alpha^t r(X_t, A_t, W_t) \right]$$

Subject to $X_{t+1} = f(X_t, A_t, W_t)$

Under $E_x^{\pi, \gamma}$:

Given $\pi = (\pi_0, \pi_1, \dots)$, $\gamma = (\gamma_0, \gamma_1, \dots)$ and start from $t = 0$, $X_t = x$

1. Simulate A_t from $\pi_t(\cdot | X_t, X_{t-1}, A_{t-1}, \dots)$
2. Simulate W_t from $\gamma_t(\cdot | X_t, A_t, X_{t-1}, A_{t-1}, \dots)$
3. Compute $X_{t+1} = f(X_t, A_t, W_t)$, $t \leftarrow t + 1$, and go back to 1.

Distributional Robustness Constraints

$$\sup_{\pi \in \Pi} \inf_{\gamma \in \Gamma} E_x^{\pi, \gamma} \left[\sum_{t=0}^{\infty} \alpha^t r(X_t, A_t, W_t) \right], \quad \text{subject to } X_{t+1} = f(X_t, A_t, W_t)$$

$$W_t \approx D_t + \delta g_t(U_t, X_t, A_t, (W_{t-1}) \dots)?$$

Given $\pi = (\pi_0, \pi_1 \dots)$, $\gamma = (\gamma_0, \gamma_1, \dots)$ and start from $t = 0$, $X_t = x$

1. Simulate A_t from $\pi_t(\cdot | X_t, X_{t-1}, A_{t-1} \dots)$
2. Simulate W_t from $\gamma_t(\cdot | X_t, A_t, X_{t-1}, A_{t-1} \dots)$
3. Compute $X_{t+1} = f(X_t, A_t, W_t)$, $t \leftarrow t + 1$, and go back to 1.

Distributional Robustness Constraints

$$\sup_{\pi \in \Pi} \inf_{\gamma \in \Gamma} E_x^{\pi, \gamma} \left[\sum_{t=0}^{\infty} \alpha^t r(X_t, A_t, W_t) \right], \quad \text{subject to } X_{t+1} = f(X_t, A_t, W_t)$$

$$W_t \approx D_t + \delta g_t(U_t, X_t, A_t, (W_{t-1}) \dots)?$$

Given $\pi = (\pi_0, \gamma)$ from $t = 0, X_t = x$

$$\{\mu \in \Delta(\mathbb{W}) : d(\mu \| \mathcal{L}(D)) \leq \delta\}$$

1. Simulate A_t from $\pi_t(\cdot | X_t, A_{t-1}, W_{t-1}, \dots) \in \mathcal{P}_\delta(D)$
2. Simulate W_t from $\gamma_t(\cdot | X_t, A_t, X_{t-1}, A_{t-1}, \dots) \in \mathcal{P}_\delta(D)$
3. Compute $X_{t+1} = f(X_t, A_t, W_t)$, $t \leftarrow t + 1$, and go back to 1.

Hence, $\Gamma = \{\gamma = (\gamma_t : t \geq 0) : \gamma_t(\cdot | \text{history}) \in \mathcal{P}_\delta(D)\}$.

Current Action Aware vs. Unaware Adversary

Current Action Awareness: $\gamma = (\gamma_t : t \geq 0) \in \Gamma$


$$W_t \sim \gamma_t(\cdot | X_t, \textcolor{red}{A}_t, X_{t-1}, A_{t-1} \dots) \in \mathcal{P}_\delta(D)$$

Current Action Unawareness: $\gamma = (\gamma_t : t \geq 0) \in \Gamma$

$$W_t \sim \gamma_t(\cdot | X_t, \textcolor{red}{a}, X_{t-1}, A_{t-1} \dots) = \mu \in \mathcal{P}_\delta(D) \text{ for every } a \in \mathbb{A}$$

Current Action Aware vs. Unaware Adversary

Current Action Awareness: $\gamma = (\gamma_t : t \geq 0) \in \Gamma$


$$W_t \sim \gamma_t(\cdot | X_t, \textcolor{red}{A}_t, X_{t-1}, A_{t-1} \dots) \in \mathcal{P}_\delta(D)$$

Adversarial environment that can respond to controller's **realized action**.

Risky, competitive environment:

System can act adversarially to controller's current action.

Current Action Aware vs. Unaware Adversary

Current Action Unawareness: $\gamma = (\gamma_t : t \geq 0) \in \Gamma$

$$W_t \sim \gamma_t(\cdot | X_t, \textcolor{red}{a}, X_{t-1}, A_{t-1} \dots) = \mu \in \mathcal{P}_\delta(D) \text{ for every } \textcolor{red}{a} \in \mathbb{A}$$

$$W_t \sim \gamma_t(\cdot | X_t, X_{t-1}, A_{t-1} \dots) \in \mathcal{P}_\delta(D)$$

$$\textcolor{blue}{X}_t = (X_{t-1} + A_{t-1} - \textcolor{blue}{W}_{t-1})_+$$

$$\text{AR}(1): W_t = D_t + \delta \textcolor{blue}{W}_{t-1},$$

W_t is dependent on $\textcolor{blue}{W}_{t-1}$, hence dependent on $\textcolor{blue}{X}_t$.

But, given history $(X_t, X_{t-1}, A_{t-1} \dots)$, W_t is independent of A_t .

Characterizing the Optimal Robust Value

$$v(x) = \sup_{\pi \in \Pi} \inf_{\gamma \in \Gamma} E_x^{\pi, \gamma} \left[\sum_{t=0}^{\infty} \alpha^t r(X_t, A_t, W_t) \right], \quad \text{s.t. } X_{t+1} = f(X_t, A_t, W_t)$$

Theorem: Assume appropriate regularity conditions.

If the adversary is current-action-**aware**, $v = u^*$ where u^* uniquely solve:

$$u^*(x) = \sup_{\phi \in \Delta(\mathbb{A})} \int_{\mathbb{A}} \inf_{\psi \in \mathcal{P}_{\delta}(D)} \int_{\mathbb{W}} r(x, a, w) + \alpha u^*(f(x, a, w)) \psi(dw) \phi(da)$$

If the adversary is current-action-**unaware**, $v = \bar{u}$ where \bar{u} uniquely solve:

$$\bar{u}(x) = \sup_{\phi \in \Delta(\mathbb{A})} \inf_{\psi \in \mathcal{P}_{\delta}(D)} \int_{\mathbb{A} \times \mathbb{W}} r(x, a, w) + \alpha \bar{u}(f(x, a, w)) \phi \times \psi(da, dw)$$

This result is an adaptation of one of the main theorems in our previous work [W et al. 23c]



Statistical complexity

Minimax Complexity for Uniform Learning

$W_t \approx \textcolor{red}{D}_t + \delta g_t(U_t, X_t, A_t, (W_{t-1}) \dots);$
 $\{\textcolor{red}{D}_t : t \geq 0\}$ are i.i.d. with **unknown** distribution.

$$\implies \mathcal{P}_\delta(\textcolor{red}{D}) := \{\mu \in \Delta(\mathbb{W}) : d(\mu \| \mathcal{L}(\textcolor{red}{D})) \leq \delta\}$$

Data set: $\textcolor{blue}{D} := \{\hat{D}_k : k = 1 \dots n\}$ i.i.d. $\hat{D}_1 \sim \mathcal{L}(\textcolor{red}{D})$

$$v(x) = \sup_{\pi \in \Pi} \inf_{\gamma \in \Gamma} E_x^{\pi, \gamma} \left[\sum_{t=0}^{\infty} \alpha^t r(X_t, A_t, W_t) \right]$$

s.t. $X_{t+1} = f(X_t, A_t, W_t)$, where f is **known**.

Learn the value function uniformly: $\sup_{x \in \mathbb{X}} |\hat{v}_{\textcolor{blue}{D}}(x) - v(x)|$

Empirical Robust Bellman Equations

Learn the value function uniformly: $\sup_{x \in \mathbb{X}} |\hat{v}_{\mathcal{D}}(x) - v(x)| < \epsilon(n)$

For the current-action-aware (CAA) case: $v = u^*$

$$u^*(x) = \sup_{\phi \in \Delta(\mathbb{A})} \int_{\mathbb{A}} \inf_{\psi \in \mathcal{P}_{\delta}(\textcolor{red}{D})} \int_{\mathbb{W}} r(x, a, w) + \alpha u^*(f(x, a, w)) \psi(dw) \phi(da)$$

For the current-action-unaware (CAU) case: $v = \bar{u}$

$$\bar{u}(x) = \sup_{\phi \in \Delta(\mathbb{A})} \inf_{\psi \in \mathcal{P}_{\delta}(\textcolor{red}{D})} \int_{\mathbb{A} \times \mathbb{W}} r(x, a, w) + \alpha \bar{u}(f(x, a, w)) \phi \times \psi(da, dw)$$

Empirical Robust Bellman Equations

Learn the value function uniformly: $\sup_{x \in \mathbb{X}} |\hat{v}_{\mathcal{D}}(x) - v(x)| < \epsilon(n)$

For the CAA: use estimator $\hat{v}_{\mathcal{D}} := u_{\mathcal{D}}^*$

$$u_{\mathcal{D}}^*(x) = \sup_{\phi \in \Delta(\mathbb{A})} \int_{\mathbb{A}} \inf_{\psi \in \mathcal{P}_{\delta}(\mathcal{D})} \int_{\mathbb{W}} r(x, a, w) + \alpha u_{\mathcal{D}}^*(f(x, a, w)) \psi(dw) \phi(da)$$

For the CAA: use estimator $\hat{v}_{\mathcal{D}} := \bar{u}_{\mathcal{D}}$

$$\bar{u}_{\mathcal{D}}(x) = \sup_{\phi \in \Delta(\mathbb{A})} \inf_{\psi \in \mathcal{P}_{\delta}(\mathcal{D})} \int_{\mathbb{A} \times \mathbb{W}} r(x, a, w) + \alpha \bar{u}_{\mathcal{D}}(f(x, a, w)) \phi \times \psi(da, dw)$$

Minimax Complexity for Uniform Learning

Theorem: If the underlying spaces $\mathbb{X}, \mathbb{A}, \mathbb{W}$, f, r, v are Lipschitz, and $k' = k/(k-1)$, then:

Θ implies that we prove a matching lower bound

Ambiguity Set $\mathcal{P}_\delta(D)$	Type	Action	Rate $\epsilon(n)$
Wasserstein	CAA ($v = u^*$)	Continuum	$\Theta(n^{-1/2})$
	CAU ($v = \bar{u}$)	Finite	
f_k -divergence	CAA ($v = u^*$)	Continuum	$\tilde{\Theta}\left(n^{-\frac{1}{k' \vee 2}}\right)$
	CAU ($v = \bar{u}$)	Finite	

Doesn't depend on $d_{\mathbb{X}}, d_{\mathbb{W}}!$

χ^2 is a special case where $k = k' = 2$



Algorithm and experimentation

Empirical Robust Bellman Equations

Learn the value function uniformly: $\sup_{x \in \mathbb{X}} |\hat{v}_{\mathcal{D}}(x) - v(x)| < \epsilon(n)$

For the CAA: use estimator $\hat{v}_{\mathcal{D}} := u_{\mathcal{D}}^*$

$$u_{\mathcal{D}}^*(x) = \sup_{\phi \in \Delta(\mathbb{A})} \int_{\mathbb{A}} \inf_{\psi \in \mathcal{P}_{\delta}(\mathcal{D})} \int_{\mathbb{W}} r(x, a, w) + \alpha u_{\mathcal{D}}^*(f(x, a, w)) \psi(dw) \phi(da)$$

For the CAA: use estimator $\hat{v}_{\mathcal{D}} := \bar{u}_{\mathcal{D}}$

$$\bar{u}_{\mathcal{D}}(x) = \sup_{\phi \in \Delta(\mathbb{A})} \inf_{\psi \in \mathcal{P}_{\delta}(\mathcal{D})} \int_{\mathbb{A} \times \mathbb{W}} r(x, a, w) + \alpha \bar{u}_{\mathcal{D}}(f(x, a, w)) \phi \times \psi(da, dw)$$

An Actor-Critic Algorithm (CAA Case)

$$\mathbf{T}_{\eta,\theta}(x) := \inf_{\psi \in \mathcal{P}_\delta(\mathcal{D})} \int_{\mathbb{W}} r(x, \pi_\eta(x), w) + \alpha u_\theta(f(x, \pi_\eta(x), w)) \psi(dw)$$

Bellman Error Minimization:  **Policy Improvement:**

$$\min_{\theta} \int_{\mathbb{X}} [u_\theta(x) - \mathbf{T}_{\eta,\theta}(x)]^2 \nu(dx) \quad \leftarrow \max_{\eta} \int_{\mathbb{X}} \mathbf{T}_{\eta,\theta}(x) \nu(dx)$$

Data Driven Inventory Control

$$X_{t+1} = (X_t + A_t - W_t)_+$$

Data: $\overline{\mathcal{D}} = \{\hat{D}_1, \dots, \hat{D}_T\}$.

Split into training $\mathcal{D} = \{\hat{D}_1, \dots, \hat{D}_n\}$ and testing $\mathcal{D}_{\text{testing}} = \{\hat{D}_{n+1}, \dots, \hat{D}_T\}$.

- Approximate $(u_{\mathcal{D}}, \pi_{\mathcal{D}})$ with (u_{θ}, π_{η}) by the actor-critic method.
- Treat the testing data as real demand $W_0 = \hat{D}_{n+1}, W_1 = \hat{D}_{n+2}, \dots$
- Evaluate π_{η} on this trajectory $\{W_t : t \geq 0\}$.

Policy Performance

$$X_{t+1} = (X_t + A_t - W_t)_+$$

Return ($\times 10^3$)	$\delta = 0.001$	$\delta = 0.01$	$\delta = 0.1$	$\delta = 0.5$	$\delta = 1$
CAA	81.72	81.83	81.97	76.48	70.65
CAU	81.46 ± 0.01	81.79 ± 0.02	82.18 ± 0.03	77.76 ± 0.03	75.04 ± 0.03
π_0^*	81.60				

Return ($\times 10^3$)	$\delta = 0.001$	$\delta = 0.01$	$\delta = 0.1$	$\delta = 0.5$	$\delta = 1$
CAA	31.75	32.21	33.17	29.67	29.64
CAU	31.46 ± 0.01	31.94 ± 0.02	33.38 ± 0.03	31.26 ± 0.2	30.48 ± 0.3
π_0^*	31.16				

Robust to Autocorrelation Within the Demand

$$X_{t+1} = (X_t + A_t - W_t)_+$$

The CAU adversary is more appropriate to handle time correlated demand:

AR(1): $W_t = D_t + \delta W_{t-1}$,

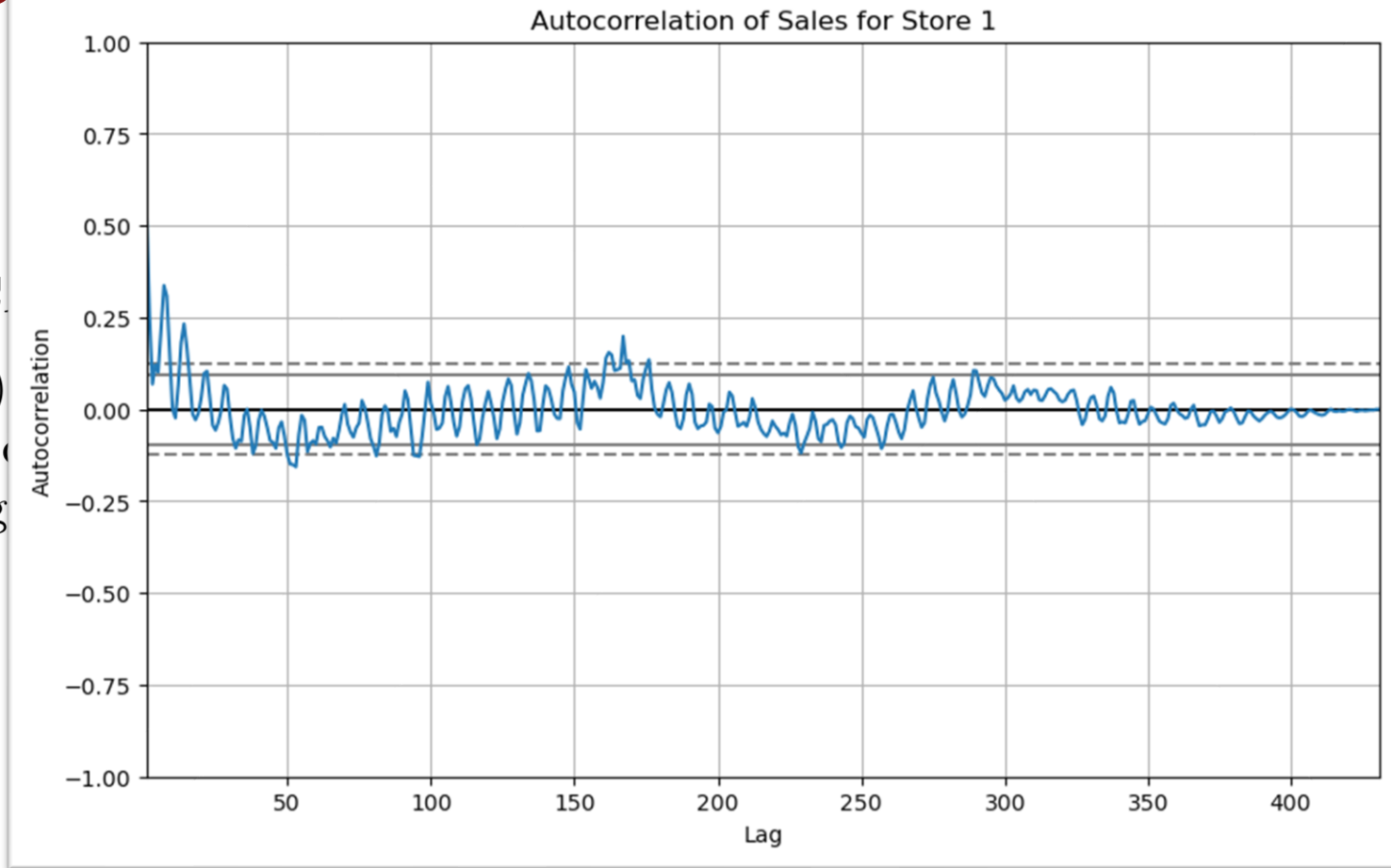
W_t is dependent on W_{t-1} , hence dependent on X_t .

But, given history $(X_t, X_{t-1}, A_{t-1} \dots)$, W_t is independent of A_t .

Robust to Autocorrelation Within the Demand

The C
AR(1)
 W_t is
But, g

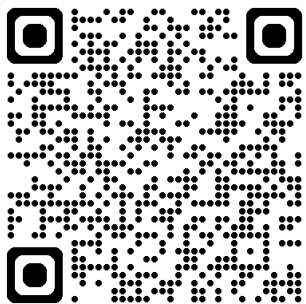
nd:





Thank You

Your questions and thoughts are most welcome!



Slides and the paper
can be found here

Summary

$v(x) = \sup_{\pi \in \Pi} \inf_{\gamma \in \Gamma} E_x^{\pi, \gamma} \sum_{t=0}^{\infty} \alpha^t r(X_t, A_t, W_t)$
s.t. $X_{t+1} = f(X_t, A_t, W_t)$, where f is **known**.

Learn the value function uniformly: $\sup_{x \in \mathbb{X}} |\hat{v}_{\mathcal{D}}(x) - v(x)| < \epsilon(n)$

For the CAA: use estimator $\hat{v}_{\mathcal{D}} := u_{\mathcal{D}}^*$

$$u_{\mathcal{D}}^*(x) = \sup_{\phi \in \Delta(\mathbb{A})} \int_{\mathbb{A}} \inf_{\psi \in \mathcal{P}_{\delta}(\mathcal{D})} \int_{\mathbb{W}} r(x, a, w) + \alpha u_{\mathcal{D}}^*(f(x, a, w)) \psi(dw) \phi(da)$$

For the CAA: use estimator $\hat{v}_{\mathcal{D}} := \bar{u}_{\mathcal{D}}$

$$\bar{u}_{\mathcal{D}}(x) = \sup_{\phi \in \Delta(\mathbb{A})} \inf_{\psi \in \mathcal{P}_{\delta}(\mathcal{D})} \int_{\mathbb{A} \times \mathbb{W}} r(x, a, w) + \alpha \bar{u}_{\mathcal{D}}(f(x, a, w)) \phi \times \psi(da, dw)$$

Ambiguity Set $\mathcal{P}_{\delta}(D)$	Type	Action	Rate $\epsilon(n)$
Wasserstein	CAA ($v = u^*$)	Continuum	$\Theta(n^{-1/2})$
	CAU ($v = \bar{u}$)	Finite	
f_k -divergence	CAA ($v = u^*$)	Continuum	$\tilde{\Theta}\left(n^{-\frac{1}{k' \vee 2}}\right)$
	CAU ($v = \bar{u}$)	Finite	

Minimax Complexity for Uniform Learning

Learn the value function uniformly *efficiently*:

$$\sup_{x \in \mathbb{X}} |\hat{v}_{\mathcal{D}}(x) - v(x)| \leq \tilde{O}_P(n^{-1/p})$$

where p doesn't depend on the dimension of \mathbb{X}, \mathbb{W} .

Hardest problem
for that estimator

Lower bound:

$$\inf_{\tilde{v}} \sup_D E^D \sup_{x \in \mathbb{X}} |\tilde{v}_{\mathcal{D}}(x) - v(x)| \geq \tilde{\Omega}(n^{-1/p}),$$

where $\mathcal{D} = \{D_1, \dots, D_n\}$ i.i.d. and $D_1 \stackrel{d}{=} D$ under E^D .

“Best” possible
estimator/algorithm