A STATISTICAL APPROACH TO APPROXIMATE DYNAMIC PROGRAMMING

Rèmi Munos[1] Csaba Szepesvári[2]

[1]Centre de Mathématiques Appliquées
Ecole Polytechnique
91128 Palaiseau Cedex, France

[2]Computer and Automation Research Inst.
of the Hungarian Academy of Sciences
Kende u. 13-17, Budapest 1111, Hungary

ICML 2006, KRL Workshop

Outline

Outline

Outline

Outline

Outline

SAMPLING BASED FITTED VALUE ITERATION – SINGLE SAMPLE

function **SFVI-SINGLE**$(N, M, K, \mu, \mathcal{F}, P, S)$

for $i = 1$ to N **do**

  Draw $X_i \sim \mu$, $Y_j^{X_i,a} \sim P(\cdot|X_i, a)$, $R_j^{X_i,a} \sim S(\cdot|X_i, a)$, $(j = 1, \ldots, M$, $a \in \mathcal{A})$

end for

$V \leftarrow 0$ // approximate value function

for $k = 1$ to K **do**

  $\hat{V}_i \leftarrow \max_{a \in \mathcal{A}} \left\{ \frac{1}{M} \sum_{j=1}^{M} \left(R_j^{X_i,a} + \gamma V(Y_j^{X_i,a}) \right) \right\}$

  $V \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{N} (f(X_i) - \hat{V}_i)^2$ // fitting

end for

return V

[Szepesvári and Munos, 2005]

SFVI is Efficient

Theorem

- MDP: smooth, stochasticity assumption satisfied.
- Fix $\mathcal{F}$, μ, ρ.
- Let π_K be greedy w.r.t. $V = \mathtt{SFVI0}(N, M, K, \mu, \mathcal{F}, P, S)$.
- Let $\epsilon = d(T\mathcal{F}, \mathcal{F})$
- With N, M, K are polynomial in the relevant quantities..
- .. with probability at least $1 - \delta$,

$$\|V^* - V^{\pi_K}\|_{p,\rho} \leq C(\mu)^{1/p} \frac{4\,\epsilon}{(1-\gamma)^2}$$

SFVI IS EFFICIENT

THEOREM

- MDP: smooth, stochasticity assumption satisfied.
- Fix $\mathcal{F}$, μ, ρ.
- Let π_K be greedy w.r.t. $V = \mathtt{SFVI0}(N, M, K, \mu, \mathcal{F}, P, S)$.
- Let $\epsilon = d(T\mathcal{F}, \mathcal{F})$
- With N, M, K are polynomial in the relevant quantities..
- .. with probability at least $1 - \delta$,

$$\|V^* - V^{\pi_K}\|_{p,\rho} \leq C(\mu)^{1/p} \frac{4\,\epsilon}{(1-\gamma)^2}$$

SFVI is Efficient

Theorem

- MDP: smooth, stochasticity assumption satisfied.
- Fix $\mathcal{F}$, μ, ρ.
- Let π_K be greedy w.r.t. $V =$SFVI0$(N, M, K, \mu, \mathcal{F}, P, S)$.
- Let $\epsilon = d(T\mathcal{F}, \mathcal{F})$
- With N, M, K are polynomial in the relevant quantities..
- .. with probability at least $1 - \delta$,

$$\|V^* - V^{\pi_K}\|_{p,\rho} \leq C(\mu)^{1/p} \frac{4\ \epsilon}{(1-\gamma)^2}$$

SFVI is Efficient

Theorem

- MDP: smooth, stochasticity assumption satisfied.
- Fix $\mathcal{F}$, μ, ρ.
- Let π_K be greedy w.r.t. $V = \texttt{SFVI0}(N, M, K, \mu, \mathcal{F}, P, S)$.
- Let $\epsilon = d(T\mathcal{F}, \mathcal{F})$
- With N, M, K are polynomial in the relevant quantities..
- .. with probability at least $1 - \delta$,

$$\|V^* - V^{\pi_K}\|_{p,\rho} \leq C(\mu)^{1/p} \frac{4\ \epsilon}{(1-\gamma)^2}$$

SFVI is Efficient

Theorem

- MDP: smooth, stochasticity assumption satisfied.
- Fix $\mathcal{F}$, μ, ρ.
- Let π_K be greedy w.r.t. $V =$SFVI0$(N, M, K, \mu, \mathcal{F}, P, S)$.
- Let $\epsilon = d(T\mathcal{F}, \mathcal{F})$
- With N, M, K are polynomial in the relevant quantities..
- .. with probability at least $1 - \delta$,

$$\|V^* - V^{\pi_K}\|_{p,\rho} \leq C(\mu)^{1/p} \frac{4\,\epsilon}{(1-\gamma)^2}$$

SFVI is Efficient

Theorem

- MDP: smooth, stochasticity assumption satisfied.
- Fix $\mathcal{F}$, μ, ρ.
- Let π_K be greedy w.r.t. $V = \texttt{SFVI0}(N, M, K, \mu, \mathcal{F}, P, S)$.
- Let $\epsilon = d(T\mathcal{F}, \mathcal{F})$
- With N, M, K are polynomial in the relevant quantities..
- .. with probability at least $1 - \delta$,

$$\|V^* - V^{\pi_K}\|_{p,\rho} \leq C(\mu)^{1/p} \frac{4\,\epsilon}{(1-\gamma)^2}$$

SFVI is Efficient

Theorem

- MDP: smooth, stochasticity assumption satisfied.
- Fix $\mathcal{F}$, μ, ρ.
- Let π_K be greedy w.r.t. $V = \texttt{SFVI0}(N, M, K, \mu, \mathcal{F}, P, S)$.
- Let $\epsilon = d(T\mathcal{F}, \mathcal{F})$
- With N, M, K are polynomial in the relevant quantities..
- .. with probability at least $1 - \delta$,

$$\|V^* - V^{\pi_K}\|_{p,\rho} \leq C(\mu)^{1/p} \frac{4\ \epsilon}{(1-\gamma)^2}$$

BELLMAN ERROR OF FUNCTION SETS

- Bound is in terms of the distance of the sets $T\mathcal{F}$, $\mathcal{F}$:

$$d(T\mathcal{F}, \mathcal{F}) \stackrel{\text{def}}{=} \sup_{V \in \mathcal{F}} \inf_{f \in \mathcal{F}} \|TV - f\|_{p,\mu}$$

- "Bellman error on $\mathcal{F}$"
- $\mathcal{F}$ should be large to make $d(T\mathcal{F}, \mathcal{F})$ small!
- if MDP is "smooth", TV is smooth for *any!* bounded V
- smooth functions can be well approximated
- $\Rightarrow$ assume MDP is smooth

Bellman error of function sets

- Bound is in terms of the distance of the sets $T\mathcal{F}$, $\mathcal{F}$:

$$d(T\mathcal{F},\mathcal{F}) \stackrel{\text{def}}{=} \sup_{V\in\mathcal{F}} \inf_{f\in\mathcal{F}} \|TV - f\|_{p,\mu}$$

- "Bellman error on $\mathcal{F}$"
- $\mathcal{F}$ should be large to make $d(T\mathcal{F},\mathcal{F})$ small!
- if MDP is "smooth", TV is smooth for *any!* bounded V
- smooth functions can be well approximated
- $\Rightarrow$ assume MDP is smooth

Bellman error of function sets

- Bound is in terms of the distance of the sets $T\mathcal{F}$, $\mathcal{F}$:

$$d(T\mathcal{F},\mathcal{F}) \stackrel{\text{def}}{=} \sup_{V\in\mathcal{F}} \inf_{f\in\mathcal{F}} \|TV - f\|_{p,\mu}$$

- "Bellman error on $\mathcal{F}$"
- $\mathcal{F}$ should be large to make $d(T\mathcal{F},\mathcal{F})$ small!
- if MDP is "smooth", TV is smooth for *any!* bounded V
- smooth functions can be well approximated
- $\Rightarrow$ assume MDP is smooth

Bellman error of function sets

- Bound is in terms of the distance of the sets $T\mathcal{F}$, $\mathcal{F}$:

$$d(T\mathcal{F},\mathcal{F}) \stackrel{\text{def}}{=} \sup_{V\in\mathcal{F}} \inf_{f\in\mathcal{F}} \|TV - f\|_{p,\mu}$$

- "Bellman error on $\mathcal{F}$"
- $\mathcal{F}$ should be large to make $d(T\mathcal{F},\mathcal{F})$ small!
- if MDP is "smooth", TV is smooth for *any!* bounded V
- smooth functions can be well approximated
- $\Rightarrow$ assume MDP is smooth

Bellman error of function sets

- Bound is in terms of the distance of the sets $T\mathcal{F}$, $\mathcal{F}$:

$$d(T\mathcal{F}, \mathcal{F}) \stackrel{\text{def}}{=} \sup_{V \in \mathcal{F}} \inf_{f \in \mathcal{F}} \|TV - f\|_{p,\mu}$$

- "Bellman error on $\mathcal{F}$"
- $\mathcal{F}$ should be large to make $d(T\mathcal{F}, \mathcal{F})$ small!
- if MDP is "smooth", TV is smooth for *any!* bounded V
- smooth functions can be well approximated
- $\Rightarrow$ assume MDP is smooth

Bellman error of function sets

- Bound is in terms of the distance of the sets $T\mathcal{F}$, $\mathcal{F}$:

$$d(T\mathcal{F},\mathcal{F}) \stackrel{\text{def}}{=} \sup_{V\in\mathcal{F}} \inf_{f\in\mathcal{F}} \|TV - f\|_{p,\mu}$$

- "Bellman error on $\mathcal{F}$"
- $\mathcal{F}$ should be large to make $d(T\mathcal{F},\mathcal{F})$ small!
- if MDP is "smooth", TV is smooth for *any!* bounded V
- smooth functions can be well approximated
- $\Rightarrow$ assume MDP is smooth

Metric entropy

- Bound depends on $\log \mathcal{N}(\mathcal{F}, N)$:

  metric entropy of $\mathcal{F}$

  (Metric-entropy measures 'capacity', similar to VC-dimension)
- Metric-entropy increases with the 'size' of $\mathcal{F}$!
- Previous slide said $\mathcal{F}$ should be big!
- How does this work out??

Metric entropy

- Bound depends on $\log \mathcal{N}(\mathcal{F}, N)$:

  metric entropy of $\mathcal{F}$

  (Metric-entropy measures 'capacity', similar to VC-dimension)
- Metric-entropy increases with the 'size' of $\mathcal{F}$!
- Previous slide said $\mathcal{F}$ should be big!
- How does this work out??

METRIC ENTROPY

- Bound depends on $\log \mathcal{N}(\mathcal{F}, N)$:

  metric entropy of $\mathcal{F}$

  (Metric-entropy measures 'capacity', similar to VC-dimension)
- Metric-entropy increases with the 'size' of $\mathcal{F}$!
- Previous slide said $\mathcal{F}$ should be big!
- How does this work out??

METRIC ENTROPY

- Bound depends on $\log \mathcal{N}(\mathcal{F}, N)$:

  metric entropy of $\mathcal{F}$

  (Metric-entropy measures 'capacity', similar to VC-dimension)
- Metric-entropy increases with the 'size' of $\mathcal{F}$!
- Previous slide said $\mathcal{F}$ should be big!
- How does this work out??

Metric entropy

- Bound depends on $\log \mathcal{N}(\mathcal{F}, N)$:

  metric entropy of $\mathcal{F}$

  (Metric-entropy measures 'capacity', similar to VC-dimension)
- Metric-entropy increases with the 'size' of $\mathcal{F}$!
- Previous slide said $\mathcal{F}$ should be big!
- How does this work out??

Polynomial sample complexity

Linear models:

$$\mathcal{F} = \{w^T\phi \mid \|w\| \leq A\}$$

- [Zhang, 2002]: $\log \mathcal{N}(\mathcal{F}, N) \sim \log(N)$
- independent of $\dim(\phi) \Rightarrow$ many 'features' do not harm!

Corollary

For smooth MDPs sample complexity is polynomial

Caveat

Smoothness is critical.

Polynomial sample complexity

Linear models:

$$\mathcal{F} = \{w^T \phi \mid \|w\| \leq A\}$$

- [Zhang, 2002]: $\log \mathcal{N}(\mathcal{F}, N) \sim \log(N)$
- independent of $\dim(\phi)$ $\Rightarrow$ many ‘features’ do not harm!

Corollary

For smooth MDPs sample complexity is polynomial

Caveat

Smoothness is critical.

Polynomial sample complexity

Linear models:

$$\mathcal{F} = \{w^T\phi \mid \|w\| \leq A\}$$

- [Zhang, 2002]: $\log \mathcal{N}(\mathcal{F}, N) \sim \log(N)$
- independent of $\dim(\phi)$ $\Rightarrow$ many 'features' do not harm!

Corollary

For smooth MDPs sample complexity is polynomial

Caveat

Smoothness is critical.

Polynomial sample complexity

Linear models:

$$\mathcal{F} = \{w^T\phi \mid \|w\| \leq A\}$$

- [Zhang, 2002]: $\log \mathcal{N}(\mathcal{F}, N) \sim \log(N)$
- independent of $\dim(\phi)$ $\Rightarrow$ many 'features' do not harm!

Corollary

For smooth MDPs sample complexity is polynomial

Caveat

Smoothness is critical.

Fixed sample-based Bellman residual criterion

Given $X_0, A_0, R_0, X_1, A_1, R_1, \ldots, X_N$:

$$
\begin{aligned}
L_{N,\pi}(Q,h) = \\
& \frac{1}{N}\sum_{t=1}^{N} w_t \Big\{ (R_t + \gamma Q(X_{t+1}, \pi(X_{t+1})) - Q(X_t, A_t))^2 \\
& \qquad - (R_t + \gamma Q(X_{t+1}, \pi(X_{t+1})) - h(X_t, A_t))^2 \Big\}
\end{aligned}
$$

$$
w_t = 1/\mu(A_t|X_t)
$$

ALGORITHM

ALGORITHM

1. Choose π_0, $i := 0$
2. While $(i \leq K)$ do:
3. Let $Q_{i+1} = \mathrm{argmin}_{Q \in \mathcal{F}^{\mathcal{A}}} \sup_{h \in \mathcal{F}^{\mathcal{A}}} L_{N,\pi_i}(Q,h)$
4. Let $\pi_{i+1}(x) = \mathrm{argmax}_{a \in \mathcal{A}} Q_{i+1}(x,a)$
5. $i := i + 1$

Main result

Theorem

[Antos et al., 2006] Under a number of assumptions.., for $x > 0$, with probability at least $(1 - K\exp(-x))$,

$$\|Q^* - Q^{\pi_K}\|_{2,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} C_{\rho,\nu}^{1/2} \left(\tilde{E}(\mathcal{F}) + E(\mathcal{F}) + S_{N,x}^{1/2}\right) + (2\gamma^K)^{1/2} R_{\max},$$

$$S_{N,x} = c_2 \frac{\left((\frac{V}{2}+1)\ln(N) + \ln(c_1) + \frac{1}{1+\kappa}\ln(\frac{bc_2^2}{4}) + x\right)^{\frac{1+\kappa}{2\kappa}}}{(b^{1/\kappa}N)^{1/2}}$$

Approximation errors

$$\|Q^* - Q^{\pi_K}\|_{2,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} C_{\rho,\nu}^{1/2} \left(\tilde{E}(\mathcal{F}) + E(\mathcal{F}) + S_{N,x}^{1/2} \right) + (2\gamma^K)^{1/2} R_{\max}$$

$$(T_Q f)(x,a) = r(x,a) + \gamma \int f(y, \mathrm{argmax}_a Q(y,a)) P(dy|x,a)$$

- $\tilde{E}(\mathcal{F})$: fixed-point approximation error of $\mathcal{F}$

$$\tilde{E}(\mathcal{F}) = \sup_{Q \in \mathcal{F}^{\mathcal{A}}} \inf_{f \in \mathcal{F}^{\mathcal{A}}} \|f - T_Q f\|_{2,\nu}$$

- $E(\mathcal{F})$: Bellman-residual of $\mathcal{F}$

$$E(\mathcal{F}) = \sup_{f, Q \in \mathcal{F}^{\mathcal{A}}} \inf_{h \in \mathcal{F}^{\mathcal{A}}} \|h - T_Q f\|_{2,\nu}$$

- ν: stationary distribution over the states, underlying the behavior policy

Distribution discrepancy constant

$$\|Q^* - Q^{\pi_K}\|_{2,\rho} \le \frac{2\gamma}{(1-\gamma)^2} C_{\rho,\nu}^{1/2} \left(\tilde{E}(\mathcal{F}) + E(\mathcal{F}) + S_{N,x}^{1/2} \right) + (2\gamma^K)^{1/2} R_{\max}$$

$$\begin{aligned} C_{\rho,\nu} &= (1-\gamma)^2 \sum_{m\ge 1} m\gamma^{m-1} c(m) \\ c(m) &= \sup_{\pi_1,\dots,\pi_m} \left\| \frac{d(\rho P^{\pi_1} P^{\pi_2} \dots P^{\pi_m})}{d\nu} \right\|_\infty \end{aligned}$$

Note

Let $C_\nu = \sup_{x,a} \|dP(\cdot|x,a)/d\nu\|_\infty$.
Then $C_{\rho,\nu} \le C_\nu$.

ESTIMATION ERROR

Bound:

$$\|Q^* - Q^{\pi_K}\|_{2,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} C_{\rho,\nu}^{1/2} \left(\tilde{E}(\mathcal{F}) + E(\mathcal{F}) + S_{N,x}^{1/2} \right) + (2\gamma^K)^{1/2} R_{\max}$$

$$S_{N,x} = c_2 \frac{\left((\frac{V}{2} + 1)\ln(N) + \ln(c_1) + \frac{1}{1+\kappa}\ln(\frac{bc_2^2}{4}) + x \right)^{\frac{1+\kappa}{2\kappa}}}{(b^{1/\kappa}N)^{1/2}}$$

Estimation error

$$S_{N,x} = c_2 \frac{\left((\frac{V}{2}+1)\ln(N) + \ln(c_1) + \frac{1}{1+\kappa}\ln(\frac{bc_2^2}{4}) + x \right)^{\frac{1+\kappa}{2\kappa}}}{(b^{1/\kappa}N)^{1/2}}$$

- $\{X_t\}_t$ is exponentially β-mixing with parameters (b,κ):

$$\beta_m \le \text{const}\exp(-bm^\kappa)$$

- $c_2 = O(R_{\max}^2/\mu_0|\mathcal{A}|) \sim R_{\max}^2$,
  $\mu_0 = \min_a \inf_x \mu(a|x)$, μ is the behavior policy
- $\ln(c_1) = O(|\mathcal{A}|^2 V_{\mathcal{F}^\times}\log|\mathcal{A}| + |\mathcal{A}|V_{\mathcal{F}^+} + V\ln(c_2))$
- V – effective dimension:

$$V = 3|\mathcal{A}|V_{\mathcal{F}^+} + |\mathcal{A}|^2 V_{\mathcal{F}^\times}$$

t-th action-value function:

$$Q_{t+1} = \operatorname*{argmin}_{Q \in \mathcal{F}^{\mathcal{A}}} \sup_{h \in \mathcal{F}^{\mathcal{A}}} L_{N,\pi_t}(Q,h)$$

Note: π_t depends on the data $\Rightarrow$ random

Fitting criterion:

$$\begin{aligned} L_{N,\pi_t}(Q,h) &= \\ &\frac{1}{N}\sum_{t=1}^{N} w_t \Big\{ \left(R_t + \gamma Q(X_{t+1},\pi_t(X_{t+1})) - Q(X_t,A_t)\right)^2 \\ &\qquad - \left(R_t + \gamma Q(X_{t+1},\pi_t(X_{t+1})) - h(X_t,A_t)\right)^2 \Big\} \end{aligned}$$

VC-Crossing dimension

Fitting criterion:

$$L_{N,\pi_t}(Q,h) = \frac{1}{N}\sum_{t=1}^{N} w_t \Big\{ \left(R_t + \gamma Q(X_{t+1},\pi_t(X_{t+1})) - Q(X_t,A_t)\right)^2 - \left(R_t + \gamma Q(X_{t+1},\pi_t(X_{t+1})) - h(X_t,A_t)\right)^2 \Big\}$$

$$\begin{aligned}
\mathcal{F}_\vee &= \{ f \,|\, f(x) = Q(x, \mathrm{argmax}_{a\in\mathcal{A}} Q'(x,a)),\ Q, Q' \in \mathcal{F}^{\mathcal{A}} \} \\
&= \{ f \,|\, f(x) = \sum_{a\in\mathcal{A}} g_a(x) \mathbb{I}_{\{\pi(x)=a\}},\ g_a \in \mathcal{F}, \pi \in \Pi_{\mathcal{F}} \} \\
\Pi_{\mathcal{F}} &= \{ \pi \,|\, \pi(x) = \mathrm{argmax}_{a\in\mathcal{A}} Q(x,a), Q \in \mathcal{F}^{\mathcal{A}} \}.
\end{aligned}$$

[Nobel, 1996]: regression trees with data dependent partitions

$\Rightarrow V_{\mathcal{F}^\times}$

VC-Crossing dimension

$$\mathcal{C}_2 = \{\{x \in \mathcal{X} : f_1(x) \geq f_2(x)\} : f_1, f_2 \in \mathcal{F}\}$$

$$V_{\mathcal{F}^\times} = V_{\mathcal{C}_2}$$

Notes:

1. $V_{\mathcal{F}^+} \leq V_{\mathcal{F}^\times}$
2. But: there exists $\mathcal{F}$ such that
   - $\mathcal{F} \subset \{f | f \text{ is monotoneous, bounded}\}$,
   - $\mathcal{F}$ is VC-major (system of level-sets have finite VC-dimension),
   - $V_{\mathcal{F}^+} < +\infty$,

   and $V_{\mathcal{F}^\times} = \infty$

Conclusions

- Connecting regression and reinforcement learning
- Continuous state space
- Single trajectory, exponential beta-mixing
- Fitted policy iteration with

  ..fixed Bellman-residual criterion
- Finite-time performance bound

  (Approximation error) + (estimation error)

- (Bound holds for sup-norm)

Future work

- Model selection, adaptivity (structural risk-minimization, penalties)
- Function set adapted to the problem ($d(T\mathcal{F},\mathcal{F}) \to$ min)
- Analysis/comparison of/with other algorithms (LSTD, AAVI, FQI)
- Continuous action space??
- Algebraic mixing
- On-line learning
- Inverse problems: $Pf = r$, $f =$?

References

Antos, A., Szepesvári, C., and Munos, R. (2006).
Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path.
In *COLT-2006*.
(to appear).

Nobel, A. (1996).
Histogram regression estimation using data-dependent partitions.
Annals of Statistics, 24(3):1084–1105.

Szepesvári, C. and Munos, R. (2005).
Finite time bounds for sampling based fitted value iteration.
In *ICML'2005*.

Zhang, T. (2002).
Covering number bounds of certain regularized linear function classes.
Journal of Machine Learning Research, 2:527–550.

Definition of $\|\cdot\|_{2,\nu}$

- $\|f\|_{2,\nu}^2 = \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\int |f(x,a)|^2 d\nu(x)$

β-MIXING

DEFINITION

Let $\{Z_t\}_{t=1,2,\ldots}$ be a stochastic process. Denote by $Z^{1:n}$ the collection $(Z_1, \ldots, Z_n)$, where we allow $n = \infty$. Let $\sigma(Z^{i:j})$ denote the sigma-algebra generated by $Z^{i:j}$ $(i \leq j)$. The m-th β-mixing coefficient of $\{Z_t\}$, β_m, is defined by

$$\beta_m = \sup_{t \geq 1} \mathbb{E}\left[\sup_{B \in \sigma(Z^{t+m:\infty})} |P(B|Z^{1:t}) - P(B)|\right].$$

A stochastic process is said to be β-mixing if $\beta_m \to 0$ as $m \to \infty$.

EXTENSION OF NOBEL'S (1996) LEMMA

Π: a family of partitions of $\mathcal{X}$, $m(\Pi)$: Cell-count of Π, $\mathcal{G}$ set of bounded ($|g| \leq K$), real-valued functions

$$\mathcal{G} \circ \Pi = \left\{ f = \sum_{A_j \in \pi} g_j \mathbb{I}_{\{A_j\}} : \pi = \{A_j\} \in \Pi, g_j \in \mathcal{G} \right\}.$$

$\phi_N(\cdot)$: $\forall \epsilon > 0$, the empirical ϵ-covering numbers of $\mathcal{G}$ on all subsets of the multiset $[x_1, \ldots, x_N]$ are majorized by $\phi_N(\epsilon)$.
Let $x^{1:N} \in \mathcal{X}^N$, $\mu_N(A) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}_{\{x_i \in A\}}$
Let

$$d(\pi, \pi') = d_{x^{1:N}}(\pi, \pi') = \mu_N(\pi \bigtriangleup \pi'), \quad \pi = \{A_j\}, \pi' = \{A'_j\} \in \Pi,$$

where

$$\pi \bigtriangleup \pi' = \{x \in \mathcal{X} : \exists j \neq j'; x \in A_j \cap A'_{j'}\} = \bigcup_{j=1}^{m(\Pi)} A_j \bigtriangleup A'_j,$$

Extension of Nobel's (1996) lemma II.

Lemma

Assume that $m(\Pi) < \infty$. *Then, for any* $\epsilon > 0$, $\alpha \in (0,1)$

$$\mathcal{N}_1(\epsilon, \mathcal{G} \circ \Pi, x^{1:N}) \leq \mathcal{N}\left(\frac{\alpha\epsilon}{2K}, \Pi, d_{x^{1:N}}\right) \phi_N((1-\alpha)\epsilon)^{m(\Pi)}.$$

The covering numbers for the composite action-value function space

Lemma

Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$, $|f| \leq K$, $x^{1:N} \in \mathcal{X}^N$, ϕ_N *as before.*

$$\mathcal{G}_2^1 = \{\mathbb{I}_{\{f_1(x) \geq f_2(x)\}} \,|\, f_1, f_2 \in \mathcal{F}\}.$$

Then $\forall \epsilon > 0$, $\alpha \in (0, 1)$,

$$\mathcal{N}(\epsilon, \mathcal{F}^L \times \mathcal{F}^L, l_{x^{1:N}}) \leq \mathcal{N}_1\left(\frac{\alpha\epsilon}{L(L-1)K}, \mathcal{G}_2^1, x^{1:N}\right)^{L(L-1)} \phi_N((1-\alpha)\epsilon)^L,$$

where

$$l_{x^{1:N}}((f, Q'), (g, \tilde{Q}')) = \frac{1}{N}\sum_{t=1}^{N} |f(x_t, \hat{\pi}(x_t; Q')) - g(x_t, \hat{\pi}(x_t; \tilde{Q}'))|.$$