

# BErMin: A Model Selection Algorithm for Reinforcement Learning Problems

Amir-massoud Farahmand and Csaba Szepesvári


[academic.SoloGen.net](http://academic.SoloGen.net)

[www.ualberta.ca/~szepesva](http://www.ualberta.ca/~szepesva)


Based on: Farahmand and Szepesvári, “**Model Selection in Reinforcement Learning**,”  
Machine Learning Journal (MLJ), Vol. 85, No. 3, pp. 299–332, 2011.

Given some interaction data from a sequential decision-making problem with a large state space, what is the best possible decision?

Not much a priori information about the problem.

 **Approach:** Value-based (estimate the optimal action-value function, then follow its greedy policy)

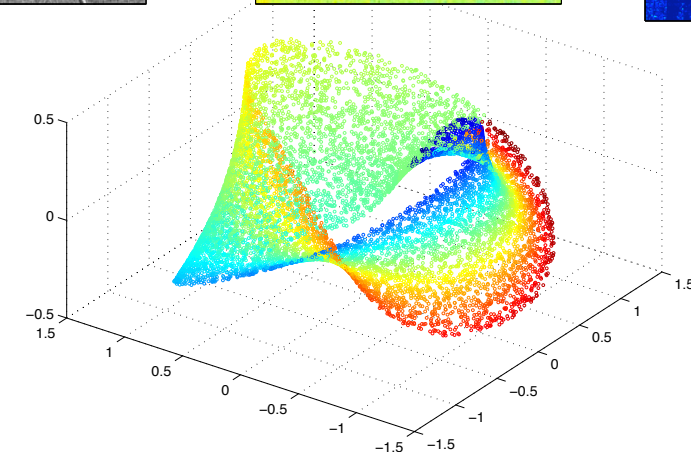
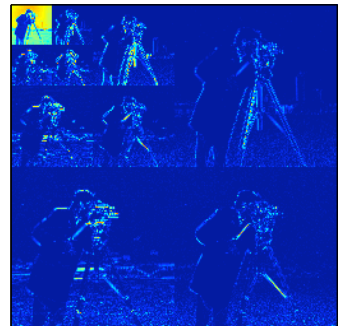
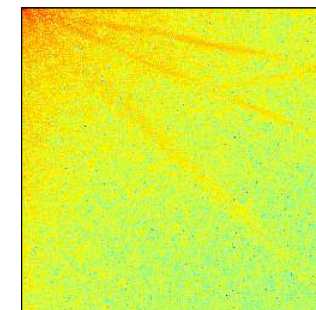
 For large state spaces, **function approximation** is required.

 **Challenge:** How to choose the architecture of the function approximator?

# How to choose the architecture of the function approximator?

- 📌 [Unknown] regularities of the value function.
- 📌 Smoothness (various notions)
- 📌 Sparsity
- 📌 Low-dimensional input manifold
- 📌 Action-gap
- 📌 Number of samples

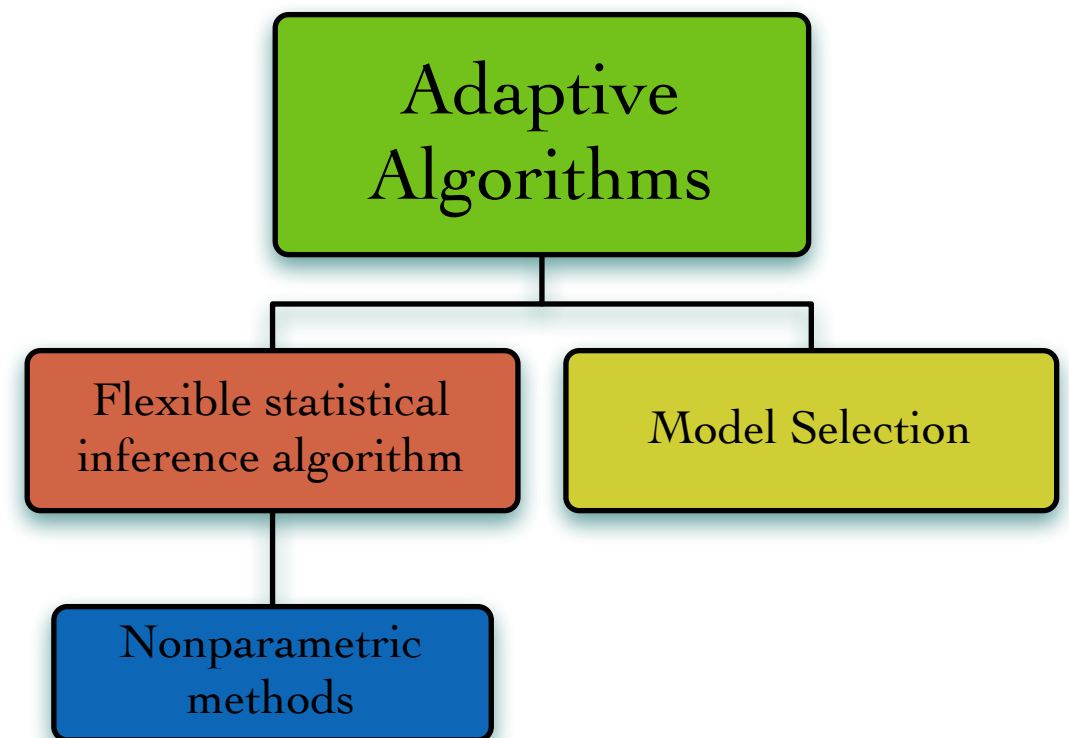
$$\int_{\mathcal{X}} |V^{(k)}(x)|^2 dx < \infty$$



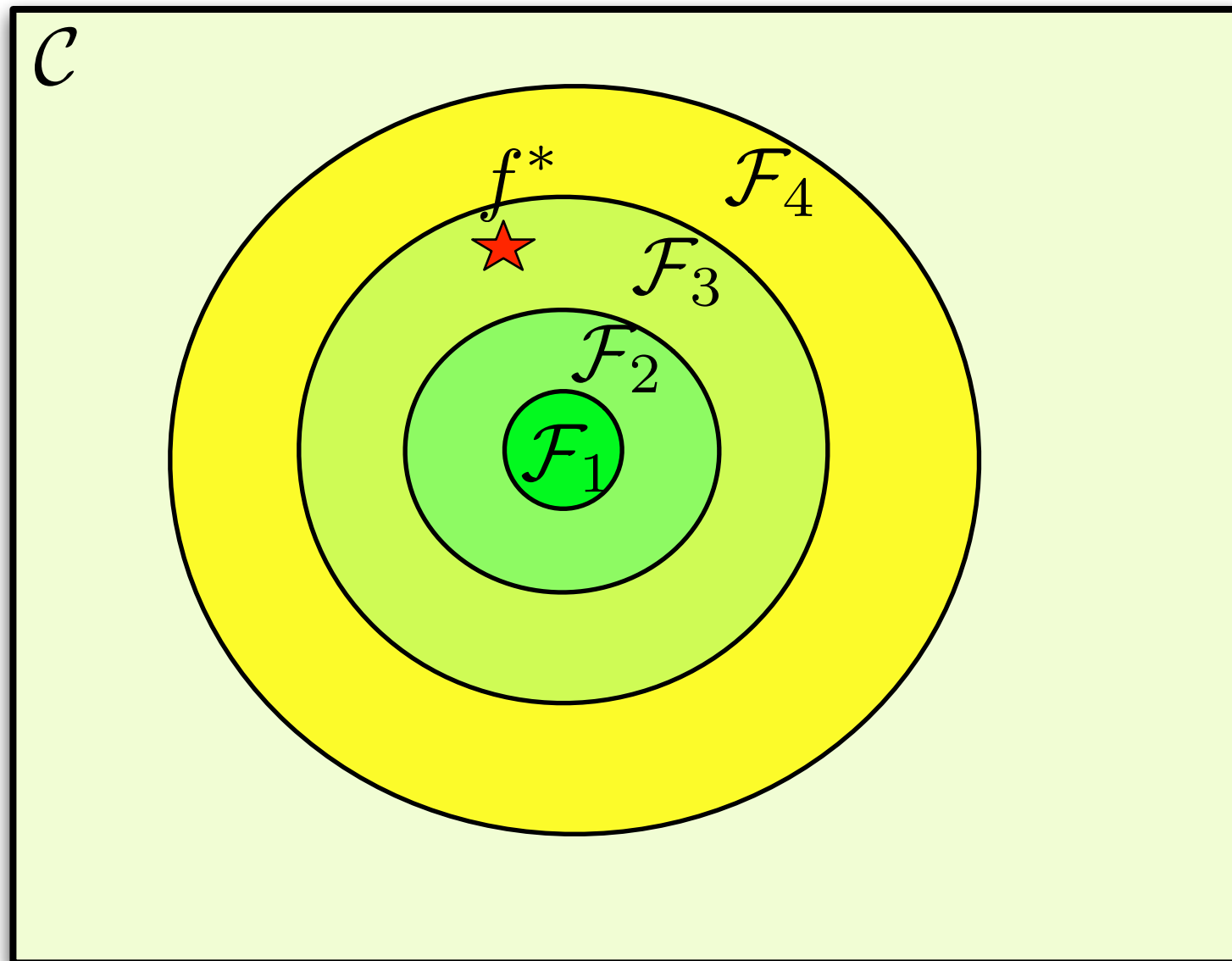
# Solution:

# Adaptive Algorithms

- **A flexible algorithm:** an algorithm that has some tunable parameters and can deliver the optimal performance for a vast range of regularities - provided that its parameters are chosen properly.
  - Examples: Regularized LSPI and Fitted Q-Iteration algorithms, Tree-based FQI, NN-based FQI, GPTD, etc.
- **A model selection algorithm:** an algorithm that tunes the parameters of a flexible algorithm.



# Regularization



$$\mathcal{F} = \bigcup_{i \geq 1} \mathcal{F}_i$$

$$\mathcal{F}_i = \{f : J(f) \leq \mu_i\} \quad (\mu_1 < \mu_2 < \cdots)$$

$J(f)$  : some measure of complexity

# Problem Setup

- Discounted MDP:  $(\mathcal{X}, \mathcal{A}, P, R, \gamma)$ .  $\mathcal{X}$  is a general state space.  $\mathcal{A}$  has finite number of actions.  $0 \leq \gamma < 1$ .
- Action-value function for policy  $\pi$ :  $Q^\pi(x, a) \triangleq \mathbb{E} \left[ \sum_{t=1}^{\infty} \gamma^{t-1} R_t \mid X_1 = x, A_1 = a \right]$
- Optimal action-value function:  $Q^*(x, a) \triangleq \sup_{\pi} Q^\pi(x, a)$ .
- Bellman optimality operator:  $(T^*Q)(x, a) \triangleq r(x, a) + \gamma \int_{\mathcal{X}} \max_{a'} Q(y, a') P(dy|x, a)$ .
- Fixed-point property:  $Q^* = T^*Q^*$ .
- Norms:  $\|Q\|_{\nu}^2 \triangleq \int_{\mathcal{X} \times \mathcal{A}} |Q(x, a)|^2 d\nu(x, a)$  and  $\|Q\|_n^2 = \frac{1}{n} \sum_{i=1}^n |Q(X_i, A_i)|^2$  for a particular set  $\{(X_i, A_i)\}_{i=1}^n$ .



**Given:** A list of action-value functions  $Q_1, Q_2, \dots, Q_P$  (with the possibility of  $P > n$ , or even  $P = \infty$ ) and a dataset

$$\mathcal{D}_n = \{(X_1, A_1, R_1, X'_1), \dots, (X_n, A_n, R_n, X'_n)\}$$

with

- $X_i \sim \nu_{\mathcal{X}}$  ( $i = 1, \dots, n$ ), with  $\nu_{\mathcal{X}}$  as the fixed distribution over the states.
- $A_i \sim \pi_b(\cdot | X_i)$  ( $\pi_b$ : data-generating policy, i.e., “*behavior*” policy).
- $R_i \sim \mathcal{R}(\cdot | X_i, A_i)$
- $X'_i \sim P(\cdot | X_i, A_i)$

**Goal:** Devise a procedure that selects the action-value function amongst  $\{Q_1, \dots, Q_P\}$  that has the smallest Bellman (optimality) error, i.e., choose  $Q_{\hat{k}}$  with

$$\hat{k} = \operatorname{argmin}_{1 \leq k \leq P} \|Q_k - T^*Q_k\|_{\nu}^2 .$$

# Challenge

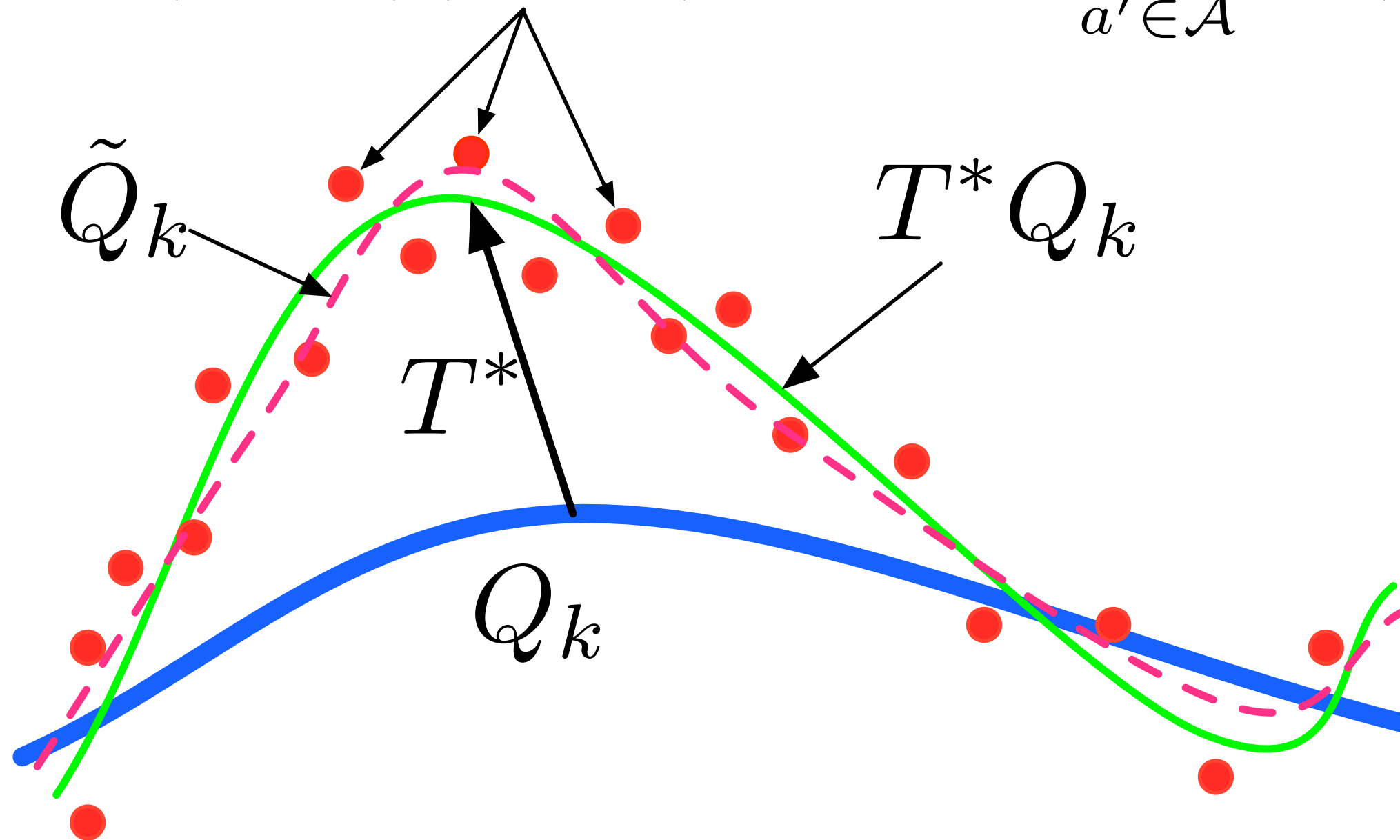
$$\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \left| Q(X_i, A_i) - [R_i + \gamma \max_{a' \in \mathcal{A}} Q(X'_i, a')] \right|^2 \right] =$$
$$\|Q - T^*Q\|_\nu^2 + \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \left| T^*Q(X_i, A_i) - [R_i + \gamma \max_{a' \in \mathcal{A}} Q(X'_i, a')] \right|^2 \right] \neq \|Q - T^*Q\|_\nu^2,$$

The variance term depends on the estimate (as opposed to supervised learning scenarios)

We **cannot** directly use empirical Bellman error to get an unbiased estimate of the true Bellman error.

What about estimating the effect of the  
Bellman operator itself?!

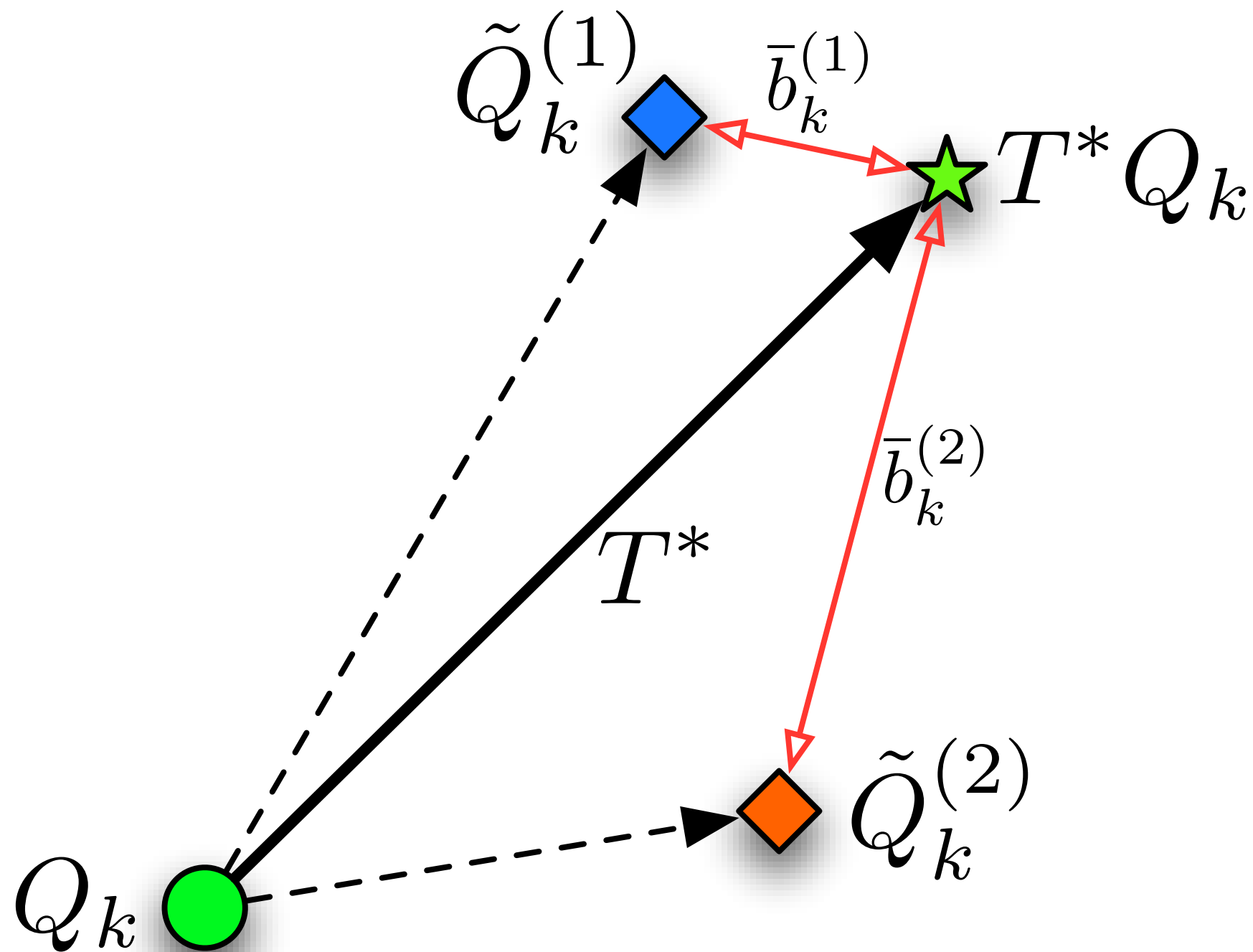
$$(\hat{T}^* Q_k)(X_i, A_i) \triangleq R_i + \gamma \max_{a' \in \mathcal{A}} Q_k(X'_i, a')$$



What if we have a good estimation of  $\tilde{Q}_k \approx T^*Q_k$  for  $k = 1, \dots, P$ ? Then we may hope that  $\|Q_k - \tilde{Q}_k\|_\nu \approx \|Q_k - T^*Q_k\|_\nu$ , and because  $\|Q_k - \tilde{Q}_k\|_\nu \approx \|Q_k - \tilde{Q}_k\|_n$  (LLN), we can use this "surrogate" risk instead.

Not done yet!

What if we have a bad estimate of the Bellman operator?



$$\frac{1}{2} \|Q_k - T^*Q_k\|_\nu^2 \leq \underbrace{\|Q_k - \tilde{Q}_k\|_\nu^2}_{\approx \|Q_k - \tilde{Q}_k\|_n^2} + \underbrace{\|T^*Q_k - \tilde{Q}_k\|_\nu^2}_{\leq \bar{b}_k \text{ (by REGRESS)}}$$

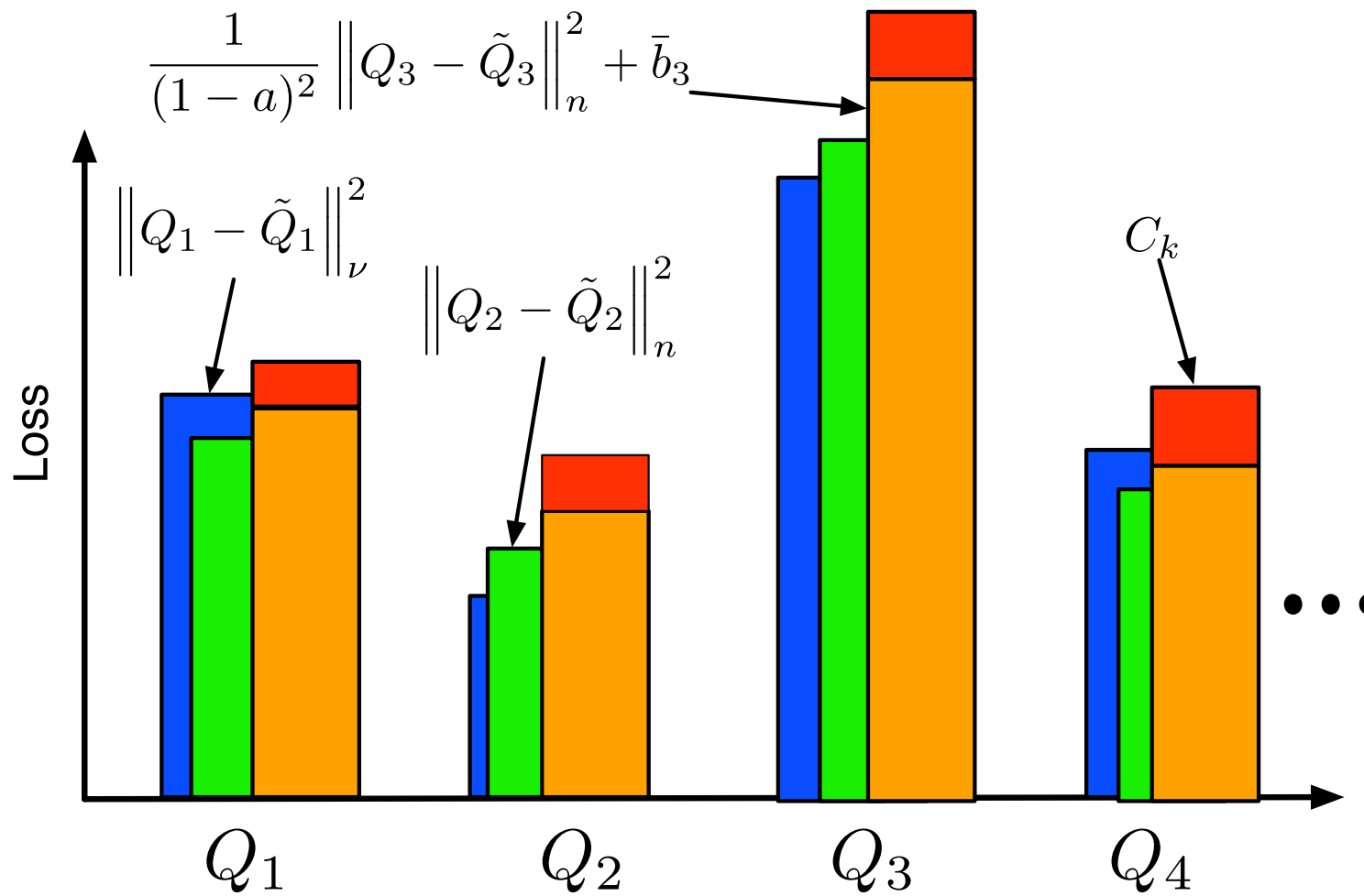
---

**Algorithm 1**  $\text{BERMIN}(\{Q_k\}_{k=1,2,\dots}, \mathcal{D}_{(m,n)}, \text{REGRESS}(\cdot), \delta, a, B, \tau)$ 

---

- 1: Split  $\mathcal{D}_{(m,n)}$  into two disjoint parts:  $\mathcal{D}_{(m,n)} = \mathcal{D}'_m \cup \mathcal{D}''_n$ .
  - 2: Choose  $(C_k)$  such that  $S = \sum_{k \geq 1} \exp(-\frac{(1-a)^2 a n}{16B^2 \tau(1+a)} C_k) < \infty$ .
  - 3: Choose  $(\delta'_k)$  such that  $\sum_{k \geq 1} \delta'_k = \delta/2$ .
  - 4: **for**  $k = 1, 2, \dots$  **do**
  - 5:      $(\tilde{Q}_k, \bar{b}_k) \leftarrow \text{REGRESS}(\mathcal{D}'_{m,k}, \delta'_k)$
  - 6:      $e_k \leftarrow \frac{1}{|\mathcal{D}''_n|} \sum_{(X,A) \in \mathcal{D}''_n} (Q_k(X, A) - \tilde{Q}_k(X, A))^2$
  - 7:      $\mathcal{R}_k^{\text{RL}} \leftarrow \frac{1}{(1-a)^2} e_k + \bar{b}_k$
  - 8: **end for**
  - 9:  $\hat{k} \leftarrow \text{argmin}_{k \geq 1} [\mathcal{R}_k^{\text{RL}} + C_k]$
  - 10: **return**  $\hat{k}$
- 

Example:  $C_k = \frac{32B^2 \tau(1+a)}{(1-a)^2 a n} \ln(k)$




---

**Algorithm 1** BERM( $\{Q_k\}_{k=1,2,\dots}, \mathcal{D}_{(m,n)}, \text{REGRESS}(\cdot), \delta, a, B, \tau$ )

---

- 1: Split  $\mathcal{D}_{(m,n)}$  into two disjoint parts:  $\mathcal{D}_{(m,n)} = \mathcal{D}'_m \cup \mathcal{D}''_n$ .
  - 2: Choose  $(C_k)$  such that  $S = \sum_{k \geq 1} \exp(-\frac{(1-a)^2 a n}{16B^2 \tau(1+a)} C_k) < \infty$ .
  - 3: Choose  $(\delta'_k)$  such that  $\sum_{k \geq 1} \delta'_k = \delta/2$ .
  - 4: **for**  $k = 1, 2, \dots$  **do**
  - 5:      $(\tilde{Q}_k, \bar{b}_k) \leftarrow \text{REGRESS}(\mathcal{D}'_{m,k}, \delta'_k)$
  - 6:      $e_k \leftarrow \frac{1}{|\mathcal{D}''_n|} \sum_{(X,A) \in \mathcal{D}''_n} (Q_k(X, A) - \tilde{Q}_k(X, A))^2$
  - 7:      $\mathcal{R}_k^{\text{RL}} \leftarrow \frac{1}{(1-a)^2} e_k + \bar{b}_k$
  - 8: **end for**
  - 9:  $\hat{k} \leftarrow \operatorname{argmin}_{k \geq 1} [\mathcal{R}_k^{\text{RL}} + C_k]$
  - 10: **return**  $\hat{k}$
-



## Assumptions:

1. The data set  $\mathcal{D}_n'' = \{(X_1, A_1, R_1, X'_1), \dots, (X_n, A_n, R_n, X'_n)\}$  is generated as described and the time-homogeneous Markov chain  $X_1, X_2, \dots, X_n$  uniformly quickly forgets its past with a forgetting time  $\tau$ .
2. The functions  $Q_k, \tilde{Q}_k, T^*Q_k$  ( $k \geq 1$ ) are bounded by a deterministic quantity  $B > 0$ .
3. The functions  $Q_k$  ( $k \geq 1$ ) are deterministic.
4. For each  $k$  and for any  $0 < \delta'_k < 1$ ,  $(\tilde{Q}_k, \bar{b}_k) = \text{REGRESS}(\mathcal{D}'_{m,k}, \delta'_k)$  are  $\sigma(\mathcal{D}'_m)$ -measurable,  $\bar{b}_k \in [0, 4B^2]$  and  $\|\tilde{Q}_k - T^*Q_k\|_\nu^2 \leq \bar{b}_k$  holds with probability at least  $1 - \delta'_k$ .
5. For  $(X_i, A_i, R_i, X'_i) \in \mathcal{D}_n''$ , the distribution of  $(X_i, A_i)$  is  $\nu$  given  $\mathcal{D}'_m$ :  $\mathbb{P}\{(X_i, A_i) \in U | \mathcal{D}'_m\} = \nu(U)$  for any measurable set  $U \subset \mathcal{X} \times \mathcal{A}$ .

**Theorem – Model Selection for RL/Planning.** Let previous assumptions hold. Consider the BERMIN algorithm used with some  $0 < a < 1$ ,  $0 < \delta \leq 1$ , and  $(C_k)_{k \geq 1}$  such that

$$S \triangleq \sum_{k \geq 1} \exp \left( -\frac{(1-a)^2 a n}{16B^2 \tau (1+a)} C_k \right) < \infty$$

holds. Let  $\hat{k}$  be the index selected by BERMIN. Then, with probability at least  $1 - \delta$ ,

$$\|Q_{\hat{k}} - T^* Q_{\hat{k}}\|_{\nu}^2 \leq 4(1+a) \min_{k \geq 1} \left\{ \frac{2}{(1-a)^2} \|Q_k - T^* Q_k\|_{\nu}^2 + \frac{3}{(1-a)^2} \bar{b}_k + 2C_k \right\} + \frac{96B^2 \tau (1+a)}{(1-a)^2 a n} \ln \left( \frac{4S}{\delta} \right) .$$

**Remember ...**

**Goal:** Devise a procedure that selects the action-value function amongst  $\{Q_1, \dots, Q_P\}$  that has the smallest Bellman (optimality) error, i.e., choose  $Q_{\hat{k}}$  with

$$\hat{k} = \operatorname{argmin}_{1 \leq k \leq P} \|Q_k - T^* Q_k\|_\nu^2 .$$

**Goal:** Devise a procedure that selects the action-value function amongst  $\{Q_1, \dots, Q_P\}$  that has the smallest Bellman (optimality) error, i.e., choose  $Q_{\hat{k}}$  with

$$\hat{k} = \operatorname{argmin}_{1 \leq k \leq P} \|Q_k - T^* Q_k\|_\nu^2.$$

**Oracle-like inequality:**

$$\|Q_{\hat{k}} - T^* Q_{\hat{k}}\|_\nu^2 \leq 4(1+a) \min_{k \geq 1} \left\{ \frac{2}{(1-a)^2} \|Q_k - T^* Q_k\|_\nu^2 + \frac{3}{(1-a)^2} \bar{b}_k + 2C_k \right\} + \frac{96B^2\tau(1+a)}{(1-a)^2 a n} \ln \left( \frac{4S}{\delta} \right)$$

$$C_k = \frac{32B^2\tau(1+a)}{(1-a)^2 a n} \ln(k)$$

# Conclusion

## What have been achieved?

- A complexity regularization-based approach for choosing a model with the minimum Bellman error.
- Oracle-like guarantee for the quality of the selected model.

## Remaining concerns:

- How to generate the list of candidates  $Q_1, \dots, Q_P$  efficiently?
- Efficient ways to estimate the excess error (i.e.,  $\bar{b}_k$ ).
- The relation of the Bellman error and the quality of the resulting policy.

**Thank you!**

Under certain assumptions, one can also prove the adaptivity.



How to estimate  $\bar{b}_k(\delta)$ ?  
The problem of excess error estimation

**Problem:** Let  $(X_1, Y_1), \dots, (X_n, Y_n)$  be a stationary, time-homogeneous Markov chain taking values in  $\mathcal{X} \times [-B, B]$  for  $\mathcal{X} \subset \mathbb{R}^d$  and let the regression function  $f^*$  be defined by  $f^*(x) = \mathbb{E}[Y_i | X_i = x]$ . Given  $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ , the goal is to provide a good estimate  $\hat{f}$  of  $f^*$  and a high confidence upper bound on the excess-risk

$$\|\hat{f} - f^*\|^2 \triangleq \|\hat{f} - f^*\|_{2,\nu}^2.$$

**Assumptions (simplified):**

- We are given a sequence of nested function spaces  $(\mathcal{F}_k)$  and  $f^* \in \cup_{k \geq 1} \mathcal{F}_k$ .
- We are given an algorithm  $A$ , which, given  $\mathcal{F}_k$ ,  $\delta$ , and a dataset of  $n$  points, returns an estimate  $\hat{f}_k$  of  $f^*$  that belongs to  $\mathcal{F}_k$ .
- For any  $k \geq 1$  there exist functions  $\mathfrak{A}_k$  and  $\mathfrak{B}_k$  such that for any  $0 < \delta \leq 1$ ,

$$L_k \triangleq \|\hat{f}_k - f^*\|^2 \leq \mathfrak{A}_k(f^*) + \mathfrak{B}_k(n, \delta, \tau)$$

holds with probability  $1 - \delta$  and that the value  $\mathfrak{B}_k(n, \delta, \tau)$ , which possibly depends on the data, can be computed at any arguments  $(n, \delta, \tau)$  and hence is available to our algorithm. No similar assumption is made about function  $\mathfrak{A}_k$ .

---

**Algorithm 2** REGRESS( $\{\mathcal{D}_n, \mathcal{D}'_n\}, \{\mathcal{F}_1, \mathcal{F}_2, \dots\}, a_n, \tau, (C_k)$ )

---

- 1: // Let  $\{(X'_t, Y'_t)\}$  be the input-output pairs in  $\mathcal{D}'_n$ :  $\mathcal{D}'_n = \{(X'_1, Y'_1), \dots, (X'_n, Y'_n)\}$ .
  - 2: **for**  $k = 1, 2, \dots$  **do**
  - 3:      $\hat{f}_k \leftarrow A(\mathcal{D}_n, \mathcal{F}_k)$ .
  - 4:      $\bar{\mathcal{R}}_k = \frac{1}{(1-a_n)^2} \frac{1}{n} \sum_{i=1}^n (\hat{f}_k(X'_i) - Y'_i)^2$ .
  - 5: **end for**
  - 6:  $\hat{k} \leftarrow \operatorname{argmin}_{k \geq 1} [\bar{\mathcal{R}}_k + C_k]$ .
  - 7: Choose  $\beta_1, \beta_2, \dots$  such that  $\beta_k \geq 0$  and  $\sum_{k \geq 1} \beta_k = 2/3$ .
  - 8: **return**  $\hat{f}_{\hat{k}}$  and  $\mathfrak{B}_{\hat{k}}(n, \cdot, \beta_{\hat{k}}, \tau)$
-

# Assumptions

Assumptions on the data:

1.  $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ ,  $\mathcal{D}'_n = \{(X'_1, Y'_1), \dots, (X'_n, Y'_n)\}$ ,  $X_i, X'_i \in \mathcal{X}$ ,  $|Y_i|, |Y'_i| \leq B$  for some  $B > 0$ .
2.  $\mathcal{D}_n$  and  $\mathcal{D}'_n$  are independent.
3.  $(X'_i, Y'_i)$  is a time-homogenous, stationary Markov chain and its forgetting time is upper bounded by  $\tau$ . We denote by  $\nu$  the stationary distribution underlying  $(X'_i)$  and we let  $\|\cdot\| = \|\cdot\|_\nu$ .

Assumptions on  $(\mathcal{F}_k)$  and the regressor function  $f^*$ :

1. The function spaces  $\mathcal{F}_1, \mathcal{F}_2, \dots$  hold measurable, real-valued functions with domain  $\mathcal{X}$  bounded by  $B > 0$ .
2. The function  $f^*(x) = \mathbb{E}[Y'_t | X'_t = x]$  belongs to  $\cup_{k \geq 1} \mathcal{F}_k$ .

Assumptions on algorithm A and functions  $\mathfrak{A}_k, \mathfrak{B}_k$ :

1. For any  $n \geq 1, k \geq 1$ , A returns a  $\sigma(\mathcal{D}_n)$ -measurable function  $\hat{f}_k$  that belongs to  $\mathcal{F}_k$  and the error bound  $L_k \triangleq \|\hat{f}_k - f^*\|^2 \leq \mathfrak{A}_k(f^*) + \mathfrak{B}_k(n, \delta, \tau)$  holds for this function with probability  $1 - \delta$ .
2. The functions  $\mathfrak{A}_k$  are such that for some  $C > 1$ ,  $\mathfrak{A}_k(f^*) \leq C \inf_{f \in \mathcal{F}_k} \|f - f^*\|^2$  holds for all  $k \geq 1$  and  $\mathfrak{A}_k(\cdot) \geq \mathfrak{A}_{k+1}(\cdot)$  holds for any  $k \geq 1$ .
3. The *known* function  $\mathfrak{B}_k(n, \delta, \tau) \xrightarrow{n \rightarrow \infty} 0$  is a decreasing function of  $n$  and an increasing function of  $\tau$ .

**Theorem – Excess Error Estimation** Assume that the conditions listed in the assumptions hold and the value of  $a_n$  given to the algorithm depends on  $n$  (e.g.,  $a_n = cn^{-1/2}$  with some  $c > 0$ ). Assume that the penalty factors,  $C_k = C_k(n)$ , passed to the excess error estimation algorithm are such that for any fixed  $k$ ,  $C_k(n)$  is a strictly decreasing function of  $n$  and for any fixed  $n$ ,

$$S_n = \sum_{k \geq 1} \exp \left( -\frac{(1 - a_n)^2 a_n n}{8B^2(1 + a_n)\tau} C_k(n) \right) < \infty.$$

Let  $\hat{f}$  and  $\hat{b}$  be the pair returned by the algorithm. Then, the following hold:  
 (A) For any  $0 < \delta \leq 1$ ,

$$\begin{aligned} \left\| \hat{f} - f^* \right\|^2 &\leq (1 - a_n^2) \min_{k \geq 1} \left[ \frac{\left\| \hat{f}_k - f^* \right\|^2}{(1 - a_n)^2} + 2C_k(n) \right] + \frac{2a_n}{1 - a_n} L(f^*) \\ &\quad + \frac{16B^2(1 + a_n)\tau \ln(\frac{2S_n}{\delta})}{(1 - a_n)a_n n} \end{aligned}$$

holds with probability at least  $1 - \delta$ , where  $L(f) = \mathbb{E} [(f(X'_1) - Y'_1)^2]$ .

(B) Fix  $0 < \delta \leq 1$ . Then, there exists  $n_0 = n_0(f^*, \delta) \geq 1$  such that for any  $n \geq n_0$ , the inequality  $\left\| \hat{f} - f^* \right\|^2 \leq \hat{b}(\delta)$  holds with probability at least  $1 - \delta$ .

Note that by selecting  $a_n \propto n^{-1/2}$ , Part (A) shows that the procedure's excess error above the oracle's performance is  $O(n^{-1/2})$ .

