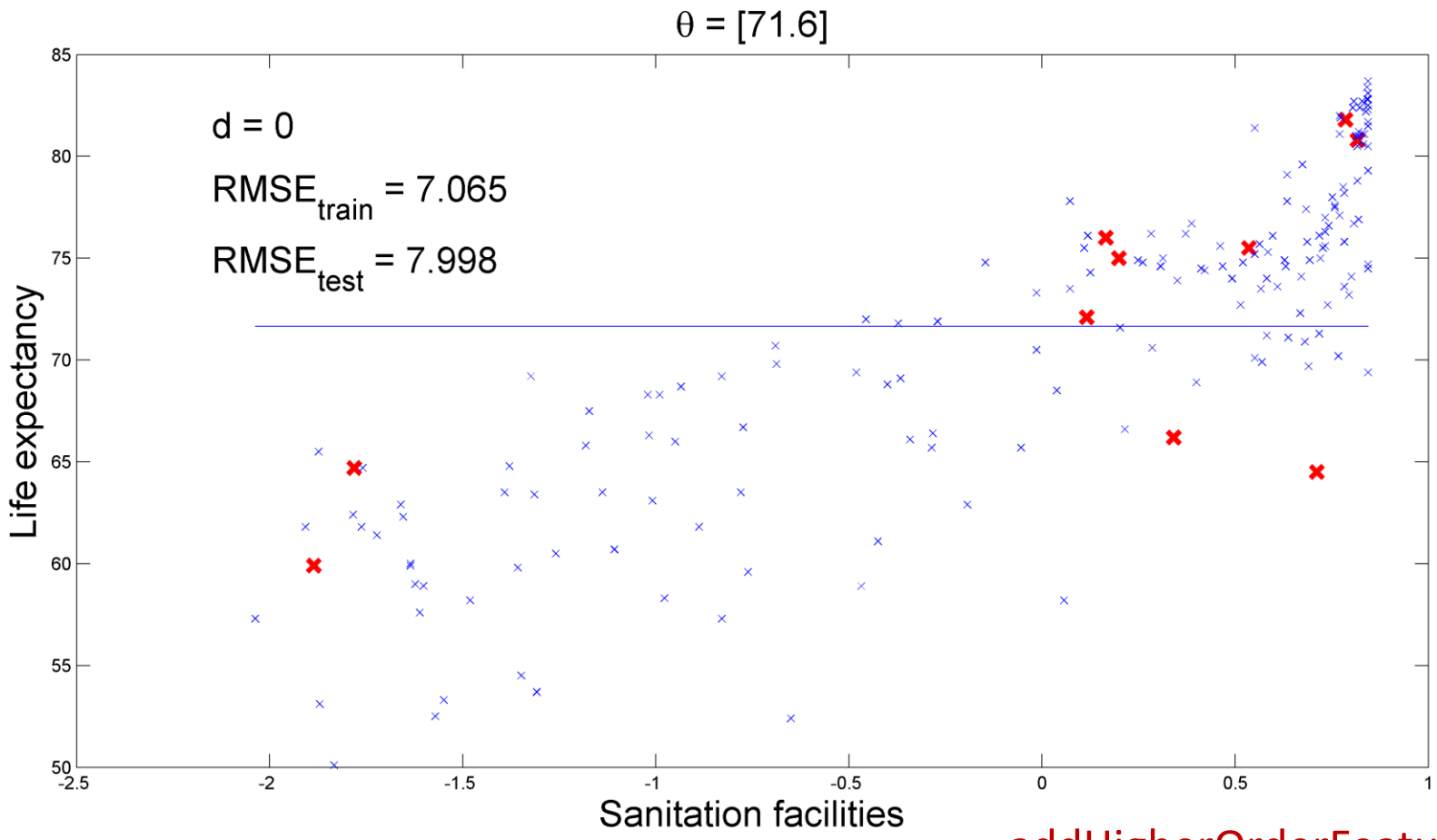
Uvođenje viših stepena polinoma

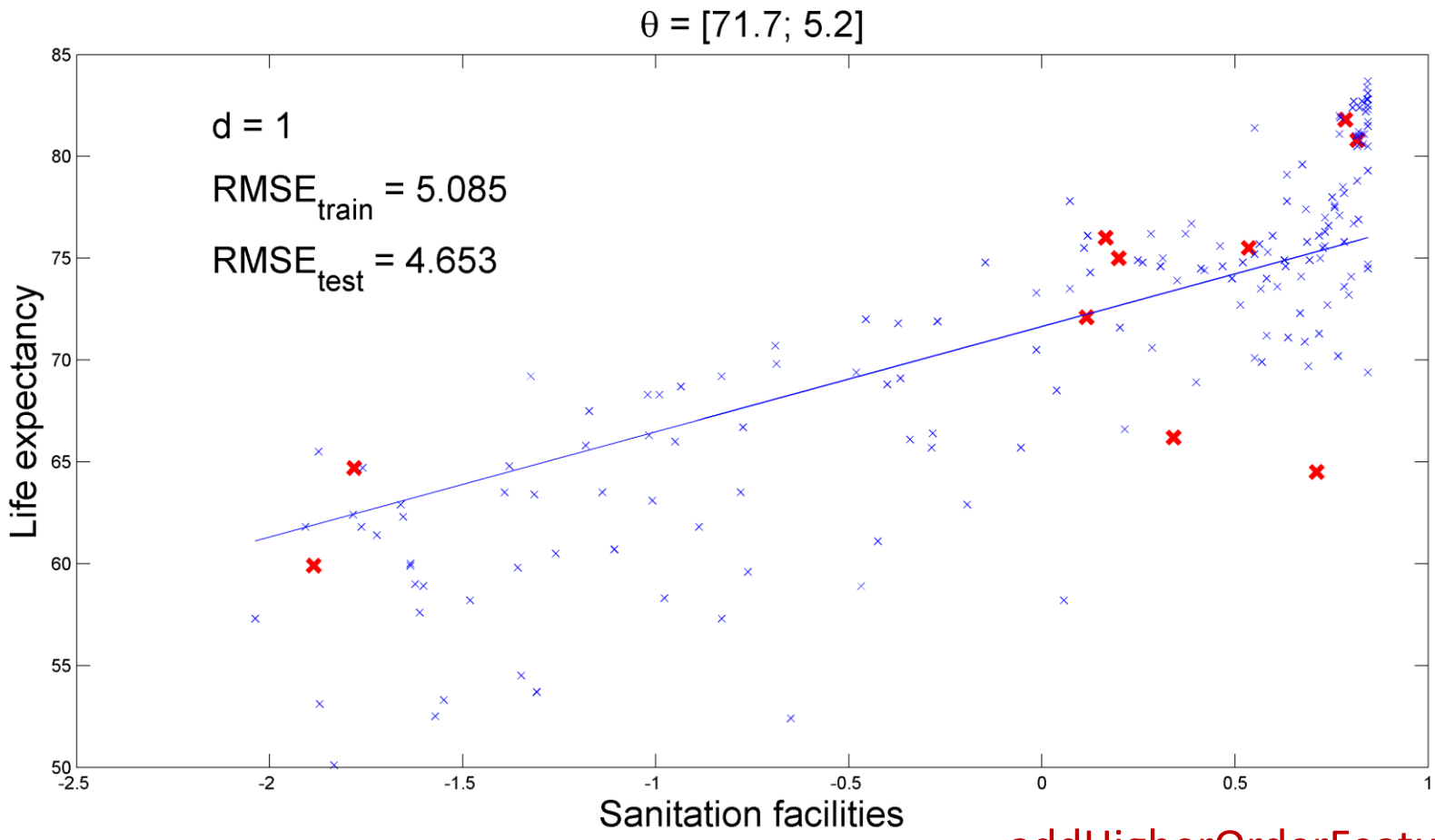
$$h_\theta(x) = \sum_{j=0}^{d} \theta_j x^j$$



addHigherOrderFeatures.m

Uvođenje viših stepena polinoma

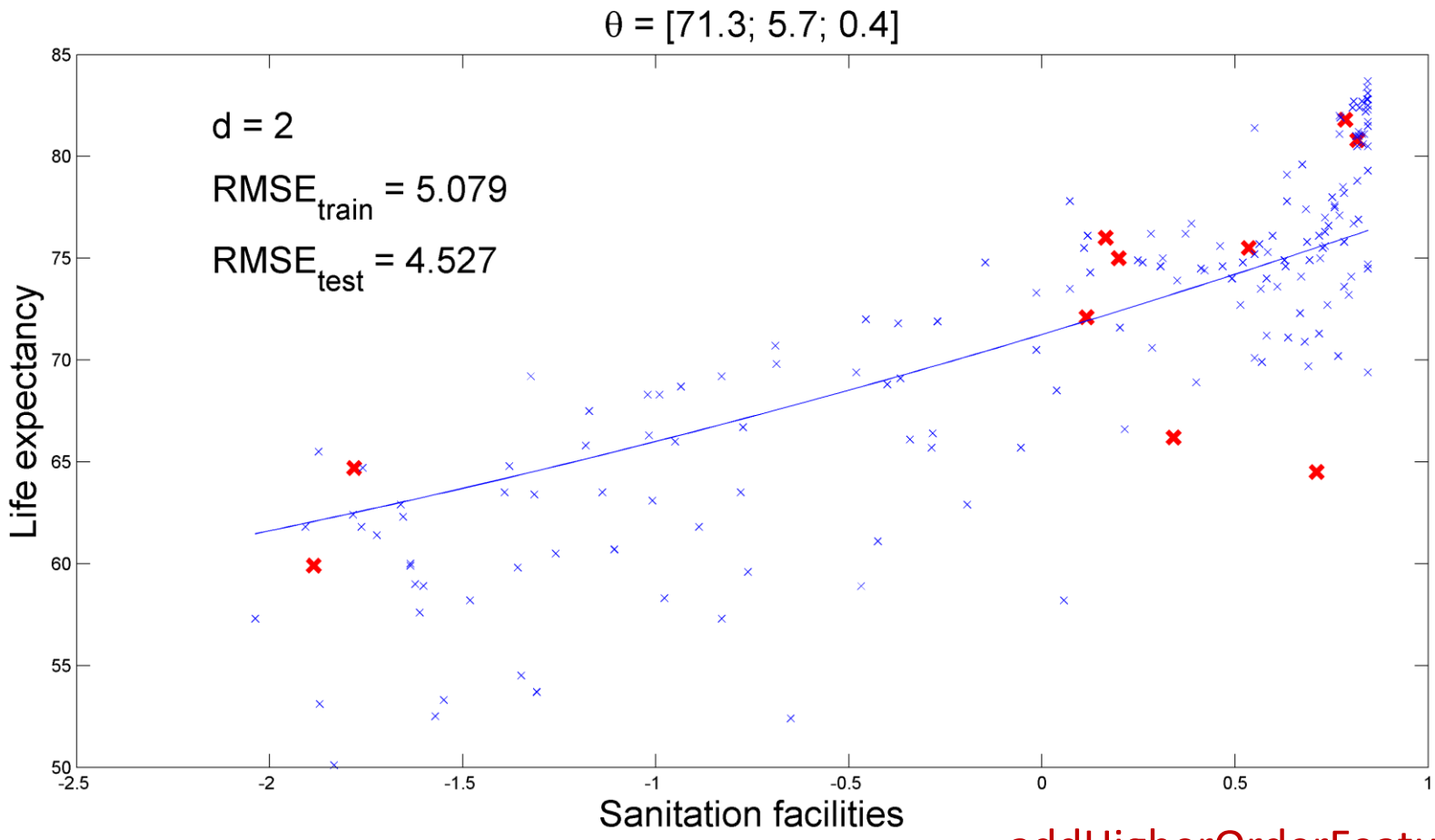
$$h_\theta(x) = \sum_{j=0}^{d} \theta_j x^j$$



addHigherOrderFeatures.m

Uvođenje viših stepena polinoma

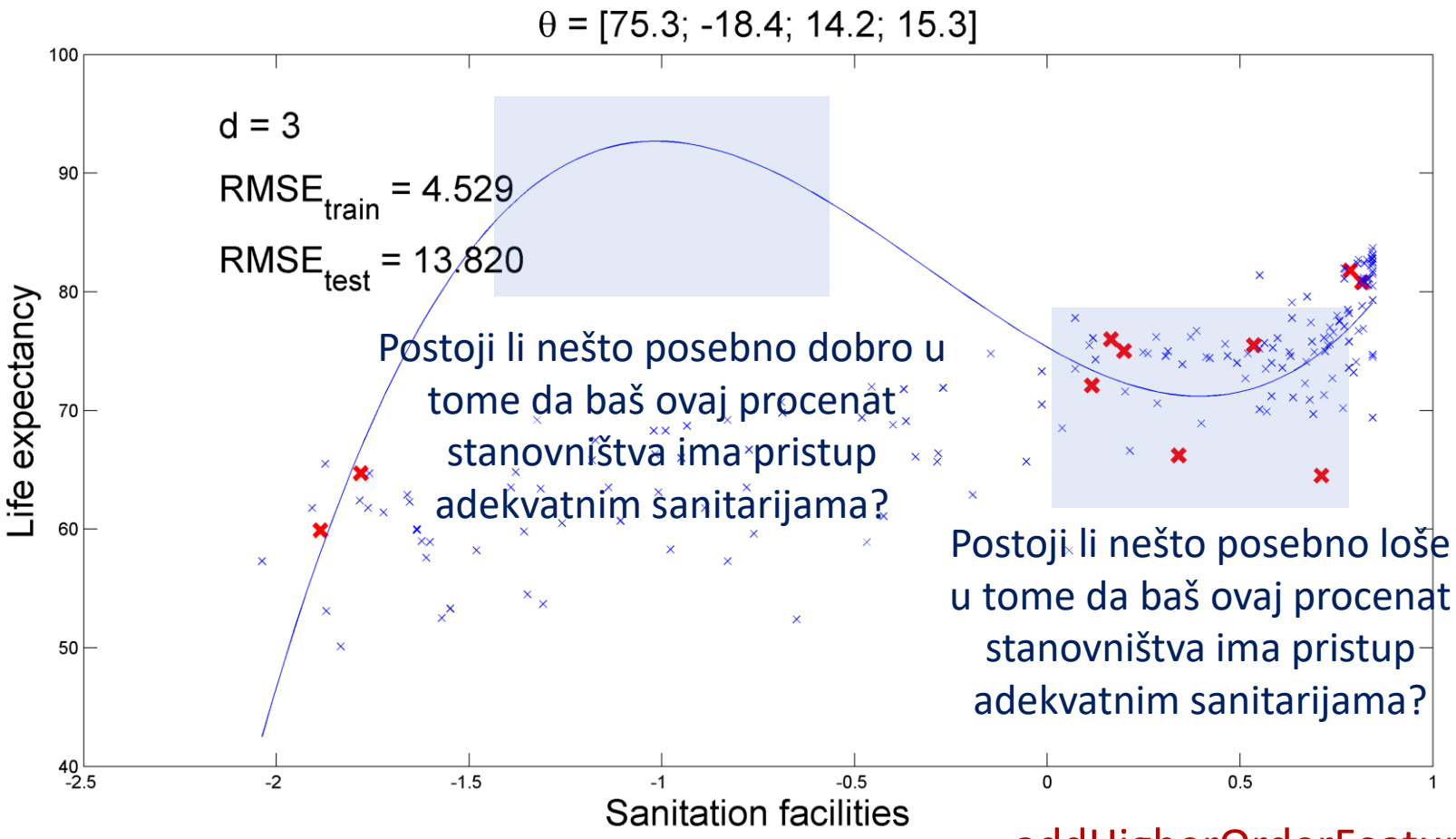
$$h_\theta(x) = \sum_{j=0}^{d} \theta_j x^j$$



addHigherOrderFeatures.m

Uvođenje viših stepena polinoma

$$h_\theta(x) = \sum_{j=0}^{d} \theta_j x^j$$

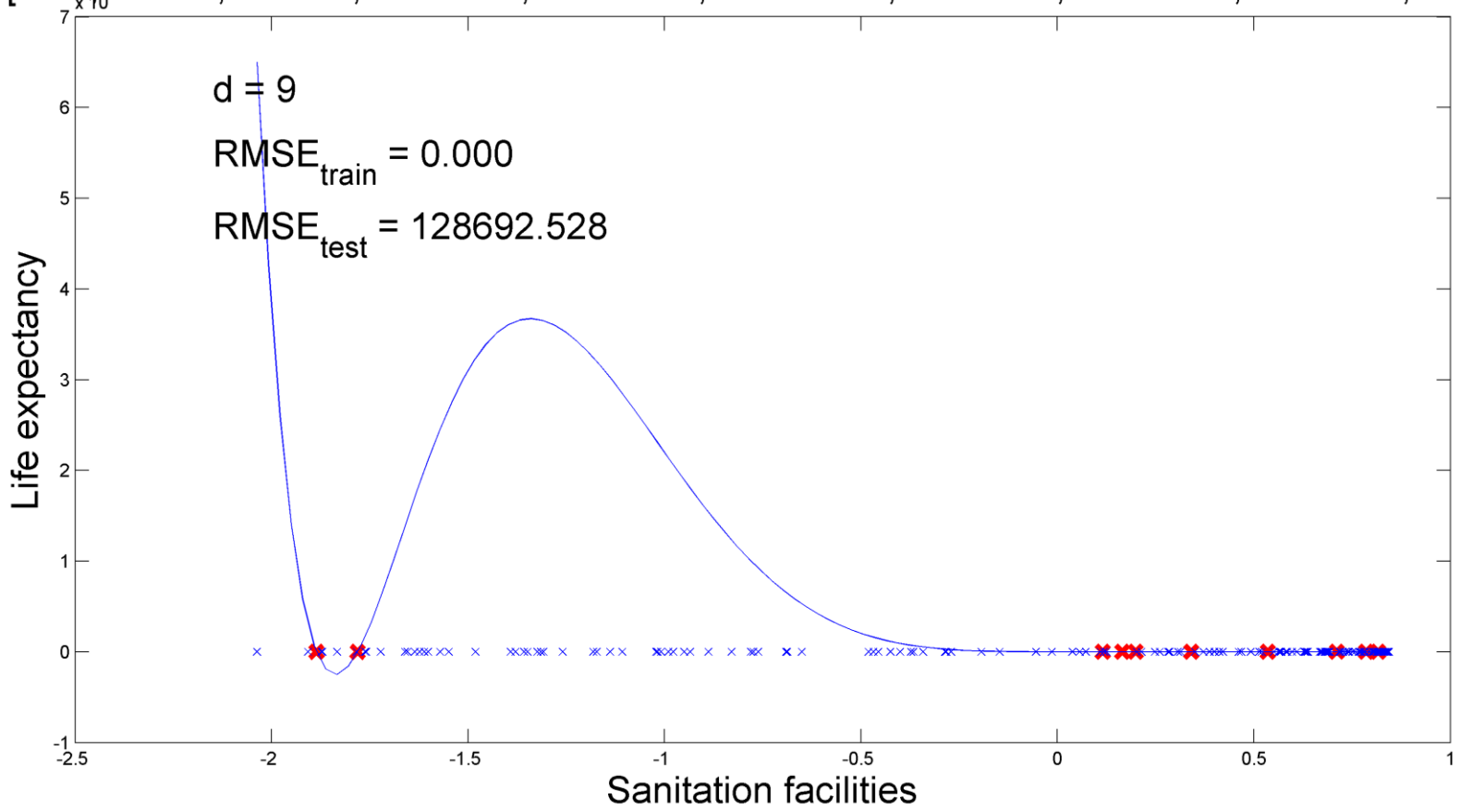


addHigherOrderFeatures.m

Uvođenje viših stepena polinoma

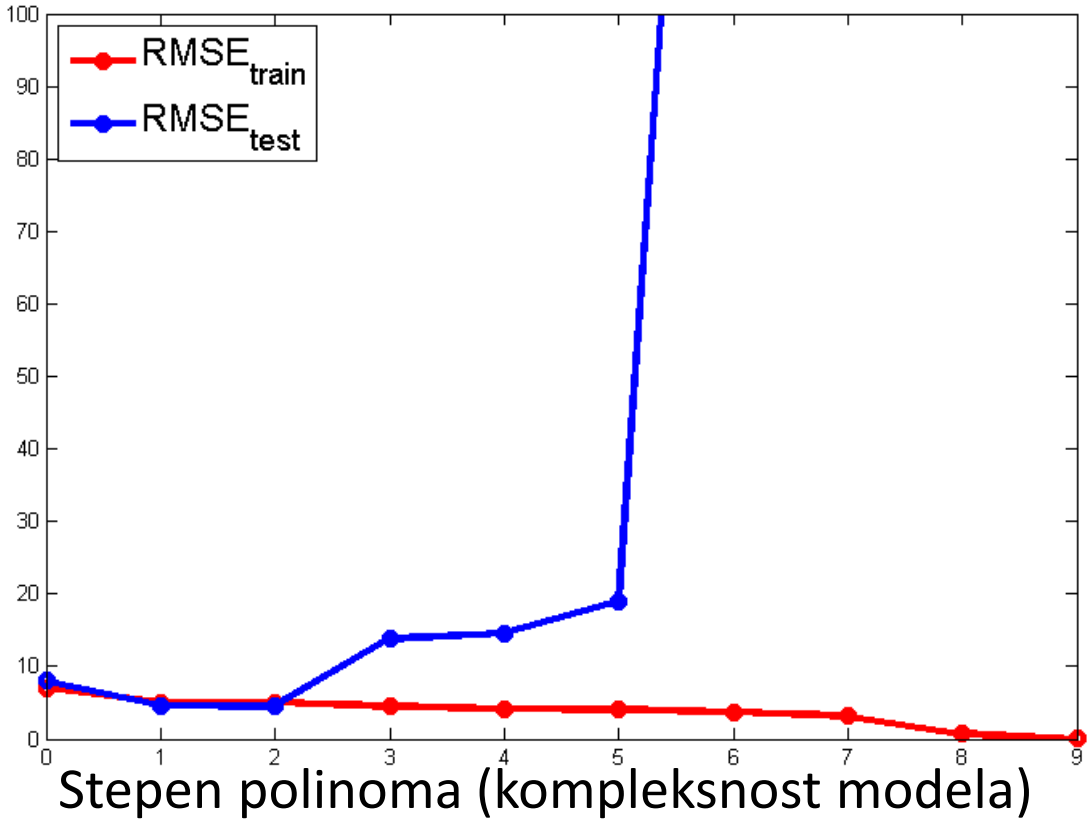
$$h_\theta(x) = \sum_{j=0}^{d} \theta_j x^j$$

θ = [54.2; -131.4; 6424.8; -48558.0; 143416.4; -161509.5; -15095.6; 132746.5; -25872.3; -33357.1]



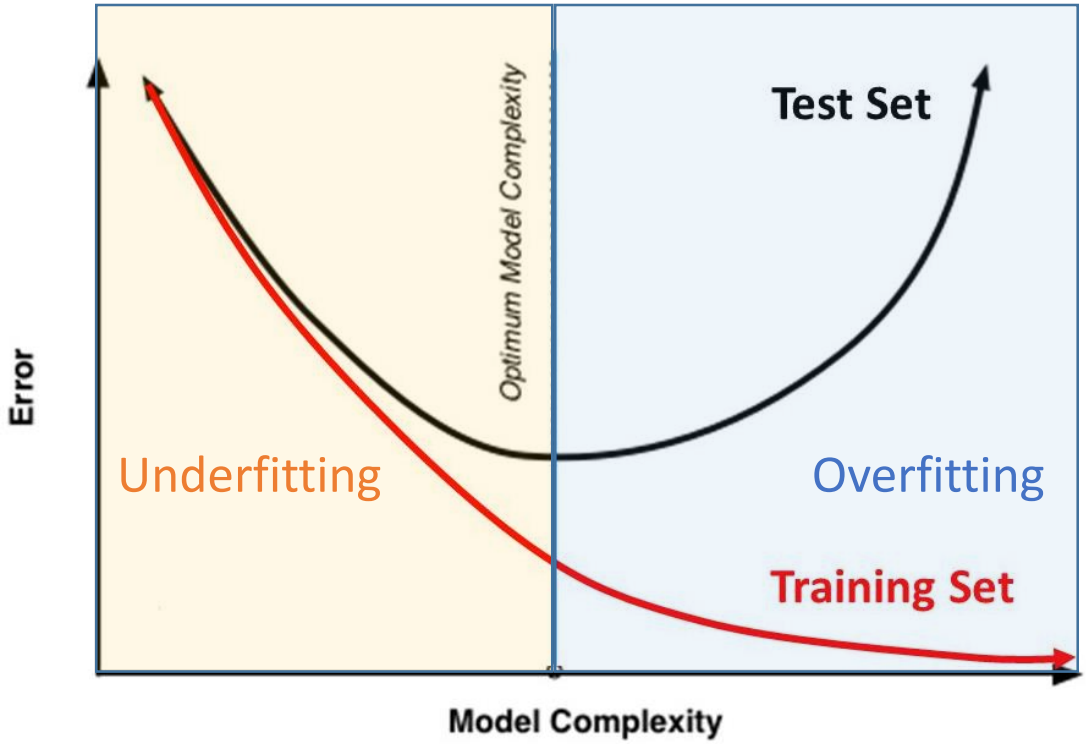
addHigherOrderFeatures.m

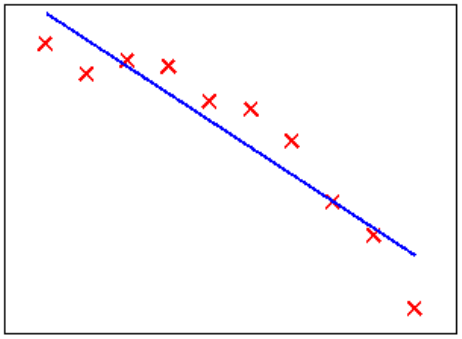
Uvođenje viših stepena polinoma



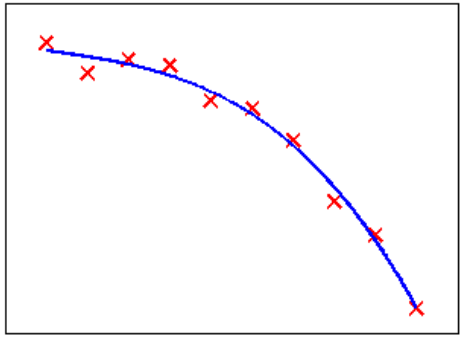
- Trening greška je preterano optimistična jer je $\hat{\theta}$ fitovan na trening podacima
- Zaključak: mala trening greška nije indikacija dobrih predikcija
  - Uporedite sa učenjem napamet

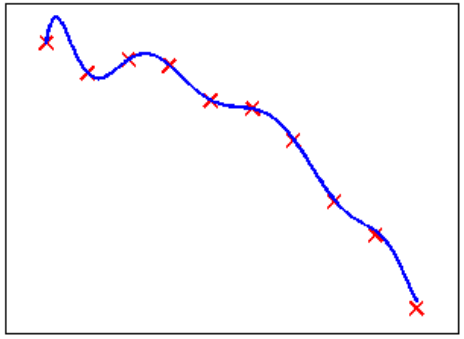
Preprilagođavanje (overfitting)





Underfit/High bias

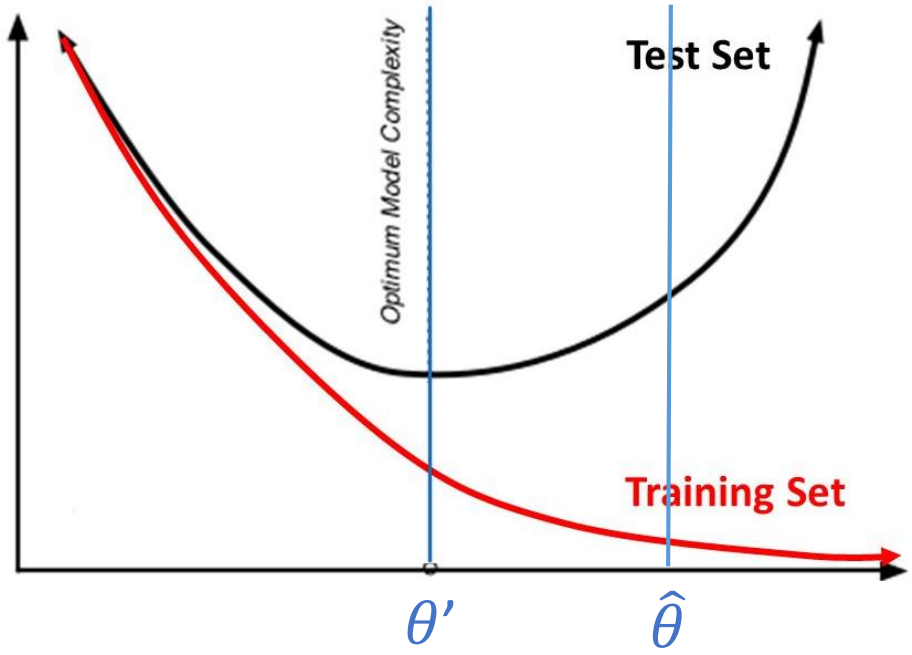




Overfit/High variance

Preprilagođavanje (overfitting)

- Previše obeležja (a premalo podataka) →
  - hipoteza se jako dobro uklapa u trening podatke
  - ali da loše generalizuje na nove primere
- Uzrok problema: prevelika prilagodljivost modela



Model $\hat{\theta}$ je overfitovan ako postoji model θ' takav da:

1) training error ($\hat{\theta}$) < training error(θ')
2) true error ($\hat{\theta}$) > true error (θ')

Preprilagođavanje (overfitting)

- Preprilagođavanje nije vezano samo za uvođenje viših stepena polinoma
- Generalno se može javiti kao posledica velikog broja obeležja (kompleksnih/fleksibilnih modela)
- Šta znači veliki broj obeležja (D) je relativno u odnosu na broj primera koji imamo u trening skupu (N)

Šta se dešava sa koeficijentima θ?

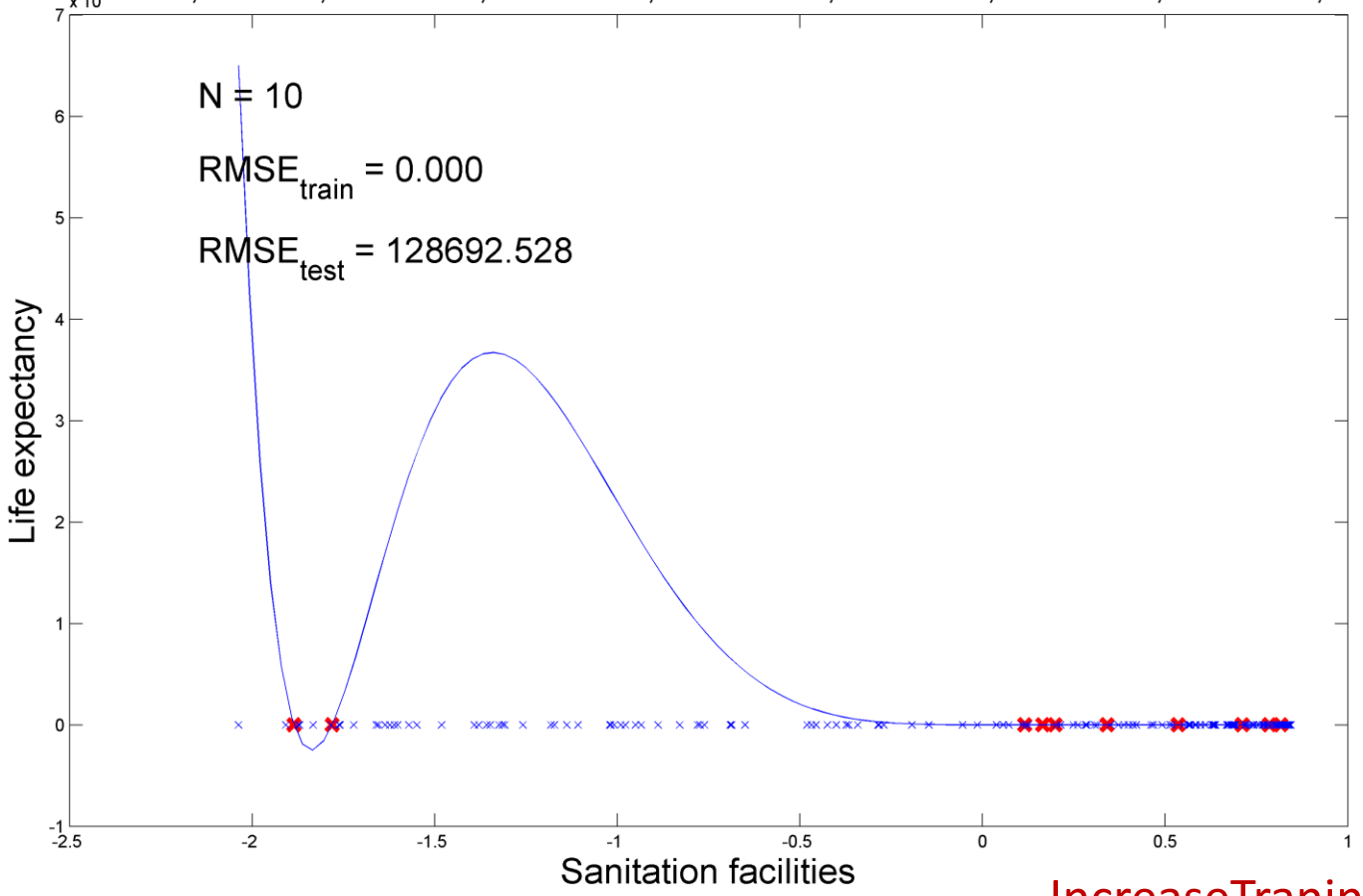
	d = 0	d = 1	d = 2	d = 3	d = 6	d = 9
θ_0	71.6	71.7	71.3	75.3	88.5	54.2
θ_1		5.2	5.7	-18.4	-171.8	-131.4
θ_2			0.4	14.2	551.2	6 424
θ_3				15.3	-471.7	-48 558
θ_4					-417.1	143 416
θ_5					405.2	-161 509
θ_6					210.4	-15 095
θ_7						132 746
θ_8						25 872
θ_9						-33 357

- Zaključak: sa uvećanjem kompleksnosti modela raste i magnituda koeficijenata θ

Povećanje *N* (fiksirana kompleksnost modela)

- Fiksiramo kompleksnost modela (u ovom primeru polinom 9. stepena)
- Postepeno ćemo uvećavati broj (slučajno selektovanih) tačaka u trening skupu

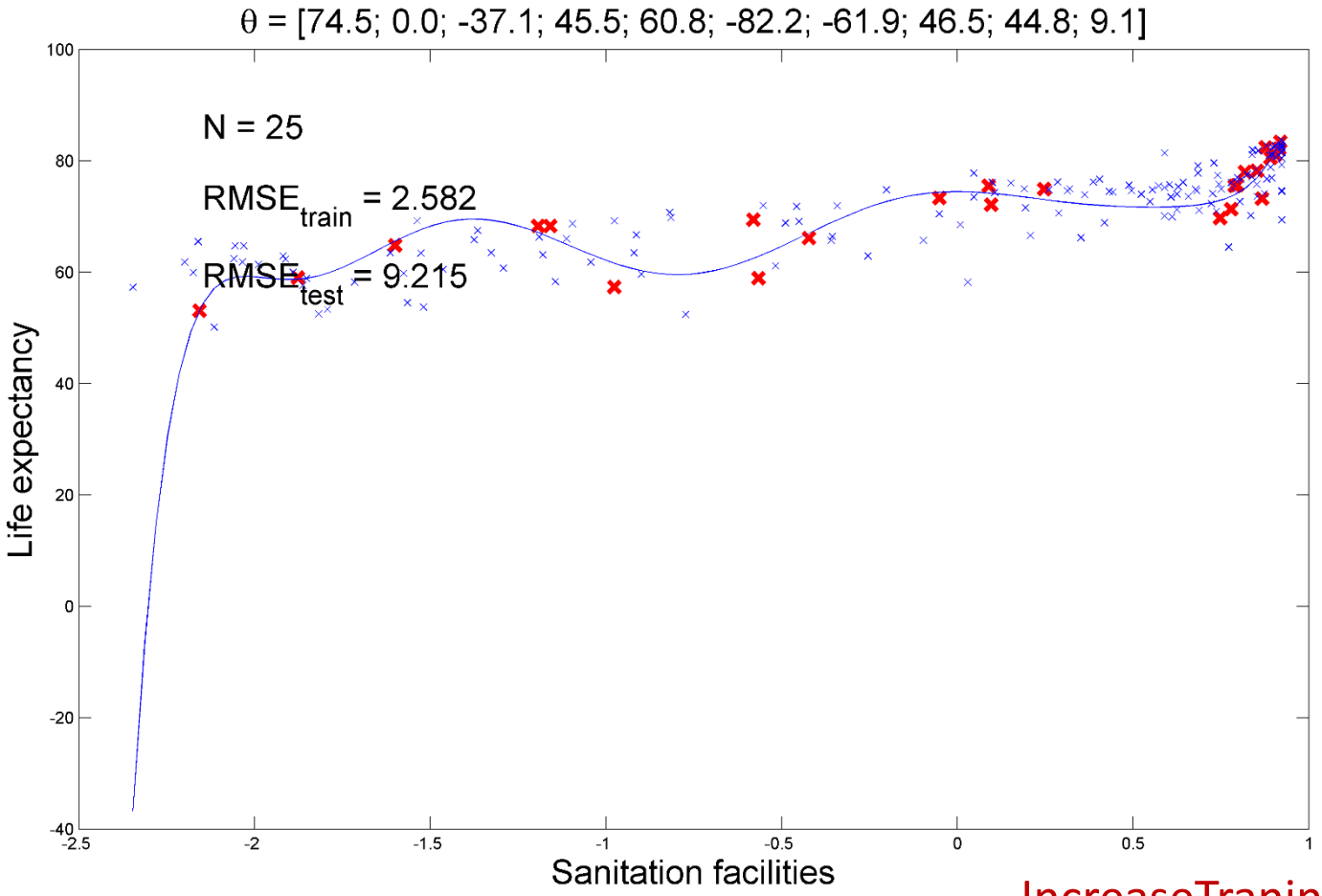
θ = [54.2; -131.4; 6424.8; -48558.0; 143416.4; -161509.5; -15095.6; 132746.5; -25872.3; -33357.1]



IncreaseTraningSetSize.m

Povećanje *N* (fiksirana kompleksnost modela)

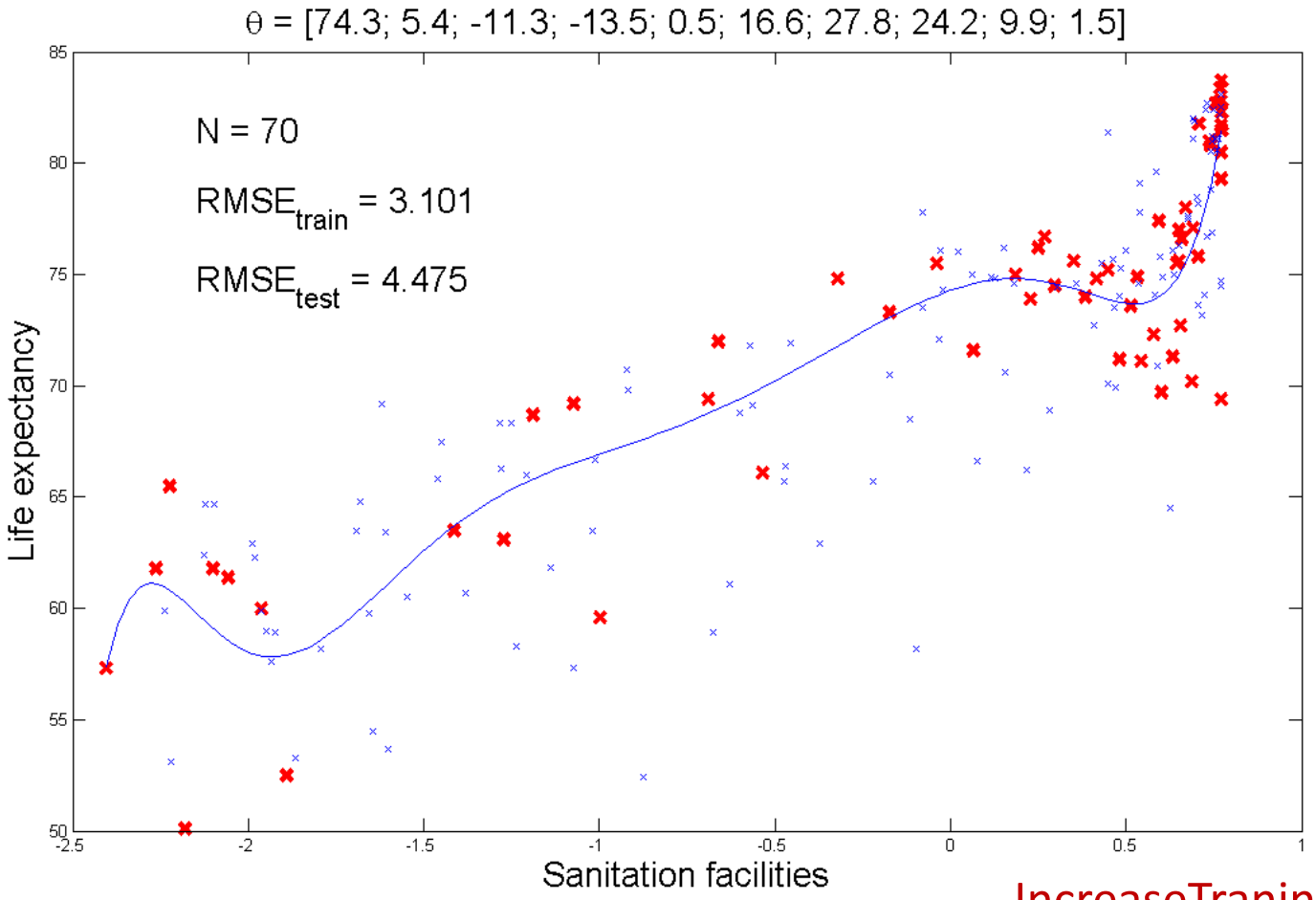
- Fiksiramo kompleksnost modela (u ovom primeru polinom 9. stepena)
- Postepeno ćemo uvećavati broj (slučajno selektovanih) tačaka u trening skupu



IncreaseTraningSetSize.m

Povećanje *N* (fiksirana kompleksnost modela)

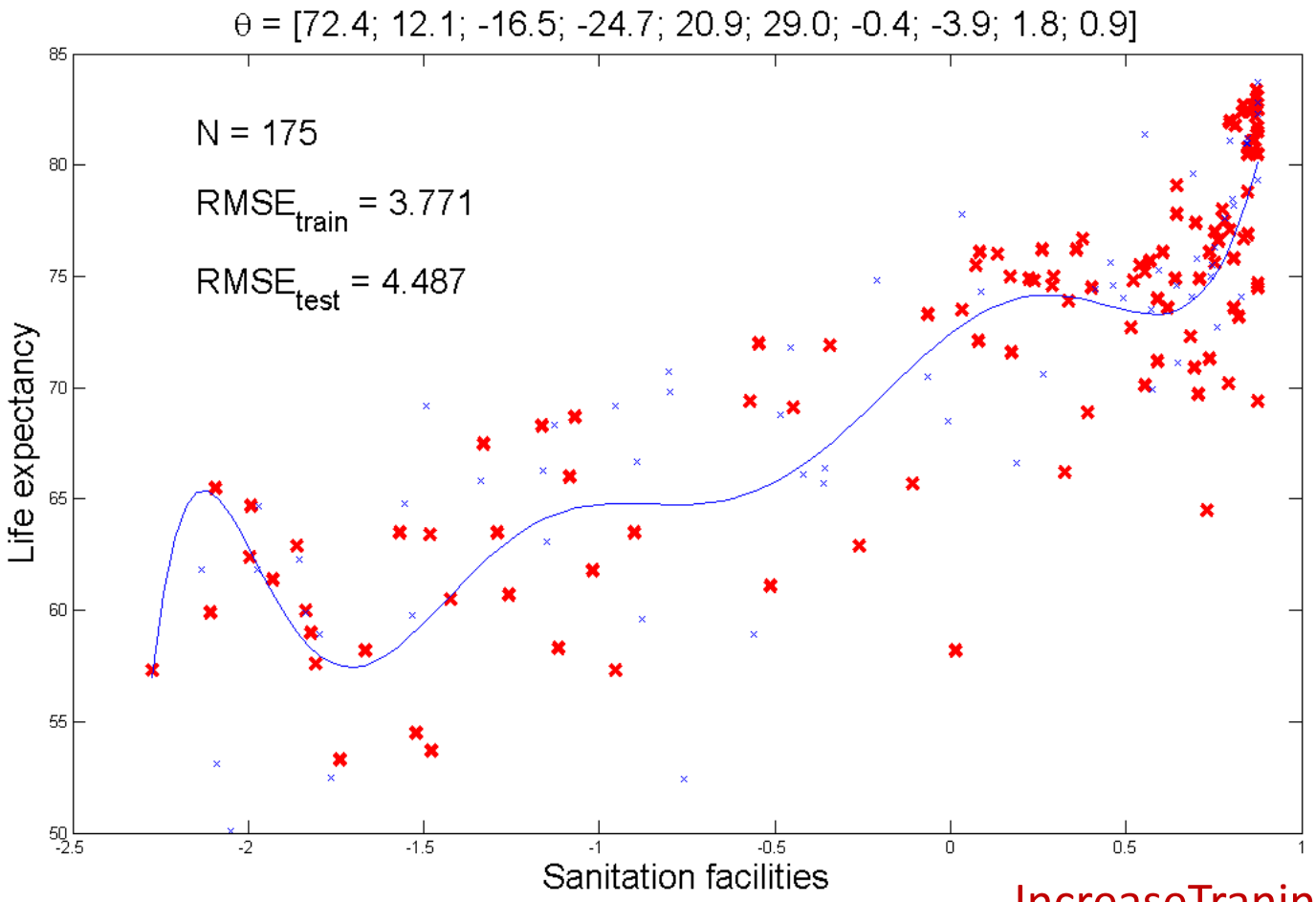
- Fiksiramo kompleksnost modela (u ovom primeru polinom 9. stepena)
- Postepeno ćemo uvećavati broj (slučajno selektovanih) tačaka u trening skupu



IncreaseTraningSetSize.m

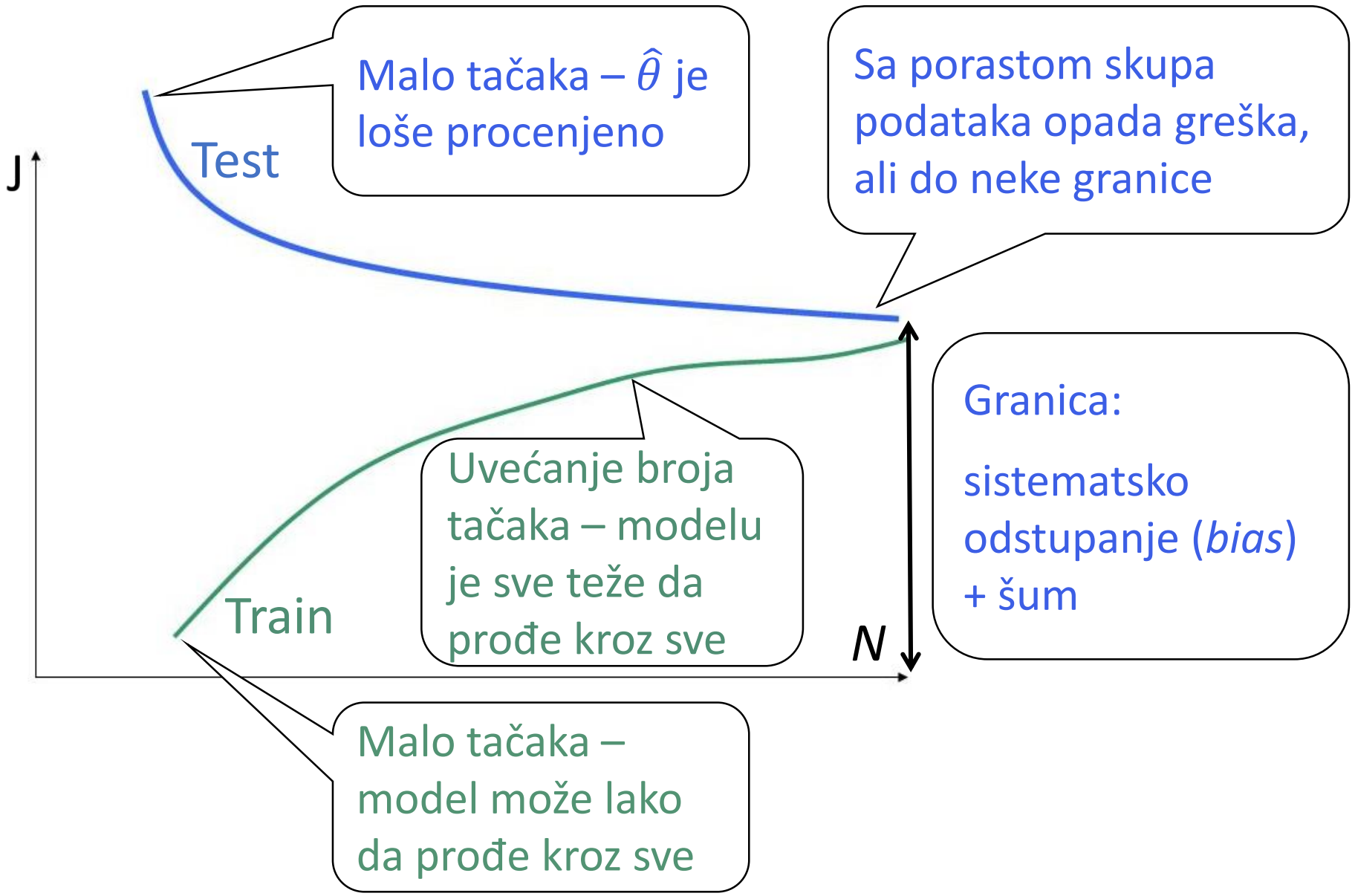
Povećanje *N* (fiksirana kompleksnost modela)

- Fiksiramo kompleksnost modela (u ovom primeru polinom 9. stepena)
- Postepeno ćemo uvećavati broj (slučajno selektovanih) tačaka u trening skupu



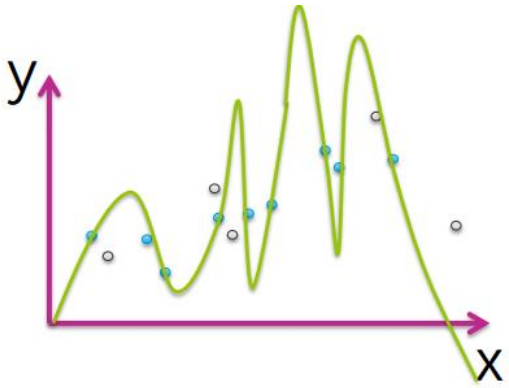
IncreaseTraningSetSize.m

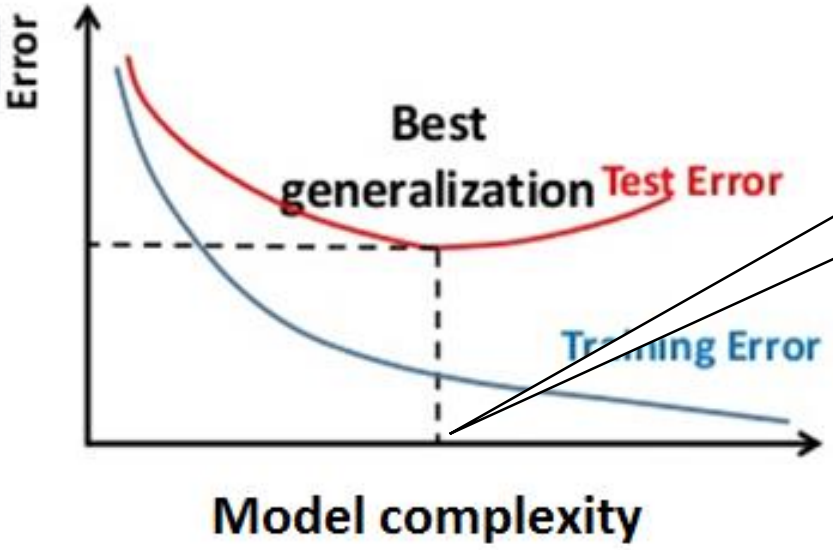
Povećanje *N* (fiksirana kompleksnost modela)



Zaključak: kako broj obeležja utiče na overfitting?

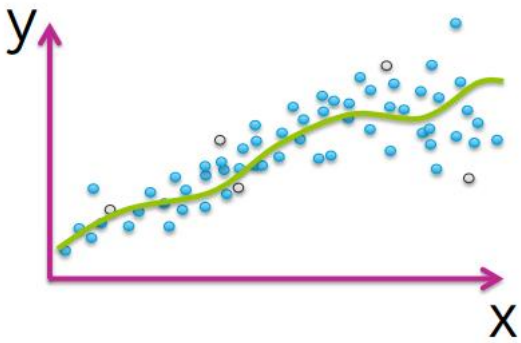
Malo primera

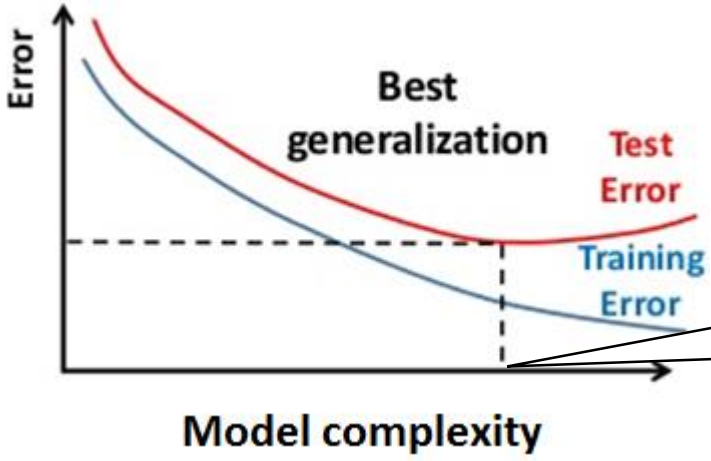




brzo dolazi do overfittinga

Mnogo primera





Teže je overfitovati

Zaključak: kako *N* utiče na preprilagođavanje?

- Zaključak: da bismo izbegli preprilagođavanje moramo imati dovoljno reprezentativan skup podataka

- Za jednu varijablu ovo je prilično teško izvesti
  - npr. da imamo primere zemalja za sve moguće odnose % stanovništva sa pristupom sanitarijama i životnog veka

- Problem se uvećava sa uvećanjem broja ulaza
  - npr. ako pored sanitarija posmatramo i % smrtnosti od saobraćajnih nesreća, teško je pokriti sve moguće kombinacije ova dva ulaza da bismo dobili reprezentativan uzorak

Zaključak: kako *N* utiče na preprilagođavanje?

- Upravljanje prilagodljivošću modela je od ključnog značaja za dobru generalizaciju

- Ovo je glavni problem mašinskog učenja i izvor njegove najdublje teorije

Šta se dešava sa koeficijentima θ?

	N = 10	N=25	N=70	N=175
θ_0	54.2	74.5	74.3	72.4
θ_1	-131.4	0.0	5.4	12.1
θ_2	6 424	-37.1	-11.3	-16.5
θ_3	-48 558	45.5	-13.5	-24.7
θ_4	143 416	60.8	0.5	20.9
θ_5	-161 509	-82.2	16.6	29.0
θ_6	-15 096	-61.9	27.8	-0.4
θ_7	132 747	46.5	24.2	-3.9
θ_8	-25 872	44.8	9.9	1.8
θ_9	-33 357	9.1	1.5	0.9