Lineáris regressziós modellek[1]

Ispány Márton és Jeszenszky Péter

2016. szeptember 19.

[1] Az ábrák C.M. Bishop: Pattern Recognition and Machine Learning c. könyvéből származnak.

Tartalom

Mintázatfelismerés

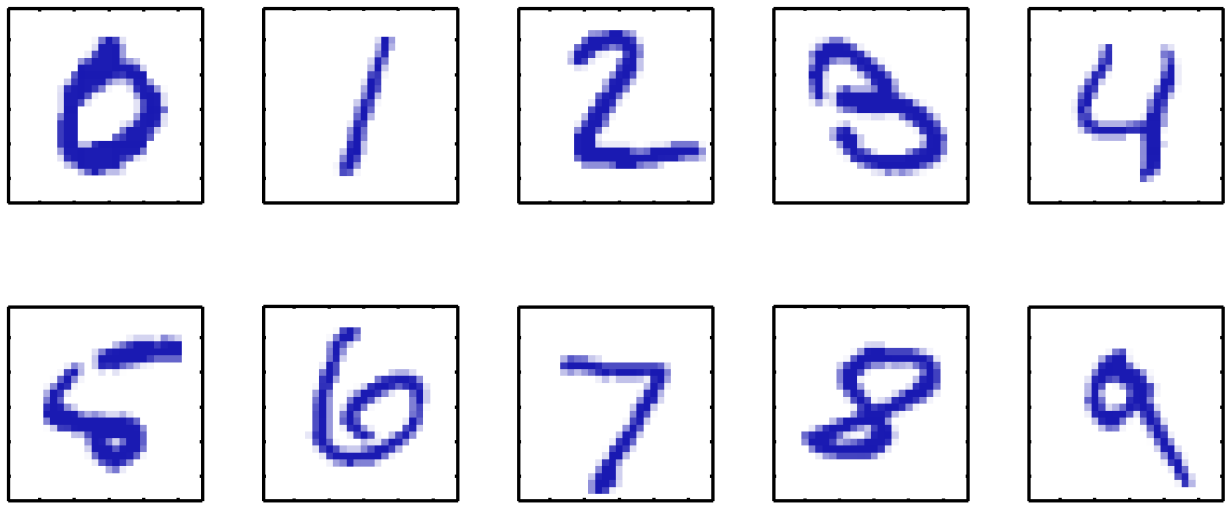
Mintázatok felismerése adatokban hosszú múltra tekint vissza.

- Tycho Brahe asztronómiai megfigyelései vezették Keplert a bolygómozgás törvényeinek felismerésére.
- Az atomi spektrumokbeli szabályosságok felfedezése alapján fejlődött ki a kvantummechanika a XX. század elején.

A fenti problémáknál a mintázat felismerését (felfedezését) emberek végezték el. A XX. század végére olyan mennyiségű adat és hozzá kapcsolódó probléma gyűlt fel, hogy ez már humán erővel kezelhetetlenné vált. Ugyanekkora tette lehetővé az informatika fejlődése, hogy ezek a problémák megoldhatóvá váltak automatikus módszerek alkalmazásával. (alakfelismerés $\rightarrow$ adatbányászat)

Karakterfelismerés (1)

Kézzel írt számjegyek felismerése: minden számjegy megfeleltethető egy 28×28 pixelű képnek, amely egy 748 dimenziós valós vektorral (input vektor) reprezentálható.
Feladat: olyan gép (szoftver) készítése, amely egy $\mathbf{x}$ input vektorhoz meghatározza a $0, 1, \ldots, 9$ számjegyek valamelyikét mint outputot. A feladat nehézsége a kézírás változatosságában van.



Karakterfelismerés (2)

Az adhoc megoldások helyett érdemes a gépi tanulás megközelítését alkalmazni. Ennek során egy nagyméretű N számjegyből álló $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ tanító adatállományt használunk egy adaptív modell paramétereinek meghatározására. Ennél az adatállománynál, pl. szakértői annotálás után, ismerjük az output vagy célváltozó értékeit: $\{t_1, \ldots, t_N\}$.

A gépi tanulás eredménye egy olyan $y(\mathbf{x})$ függvényben foglalható össze, amely egy új $\mathbf{x}$ digitális képhez generál egy y célértéket. Ezen függvény pontos alakja a tanulási fázis után határozódik meg.

A számítások sokszor felgyorsíthatóak előfeldolgozás révén, ún. jellemző kinyerés alkalmazásával, ami az eredeti input változók alkalmas transzformációját jelenti.

A példa a felügyelt tanítás egy speciális esete, egy osztályozási feladat (diszkrét a célváltozó). Ha a célváltozó folytonos, akkor regressziós feladatról beszélünk.

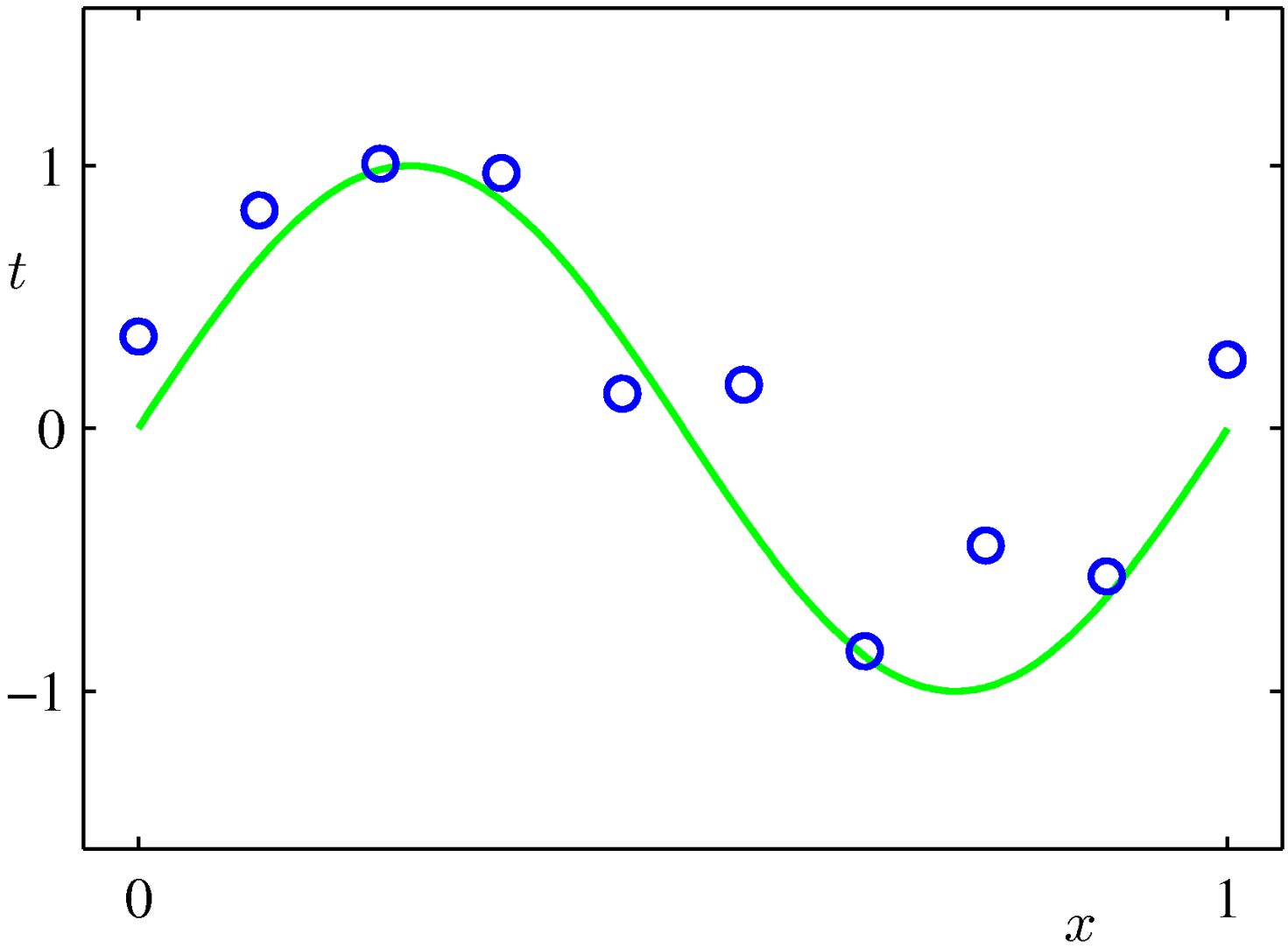
Példa: polinom illesztése

Mesterséges példa az (x, t) függvénykapcsolat gépi tanulással való vizsgálatára. Legyen az ismert kapcsolat $t = sin(2\pi x)$. Tekintsünk egy N elemű tanító állományt $\mathbf{x} = (x_1, \ldots, x_N)$ és $\mathbf{t} = (t_1, \ldots, t_N)$ értékekkel.
Legyen a közelítő függvény polinom alakú, legfeljebb M fokszámú, amelyet a következő alakban írhatunk fel

$$y(x, \mathbf{w}) := \sum_{j=0}^{M} w_j x^j$$

ahol $\mathbf{w} := (w_0, w_1, \ldots, w_M)^\top$ az ismeretlen paraméterek vagy súlyok, amelyeket a tanító adatállomány segítségével szeretnénk meghatározni.

A szinusz görbe és 10 tanító pont



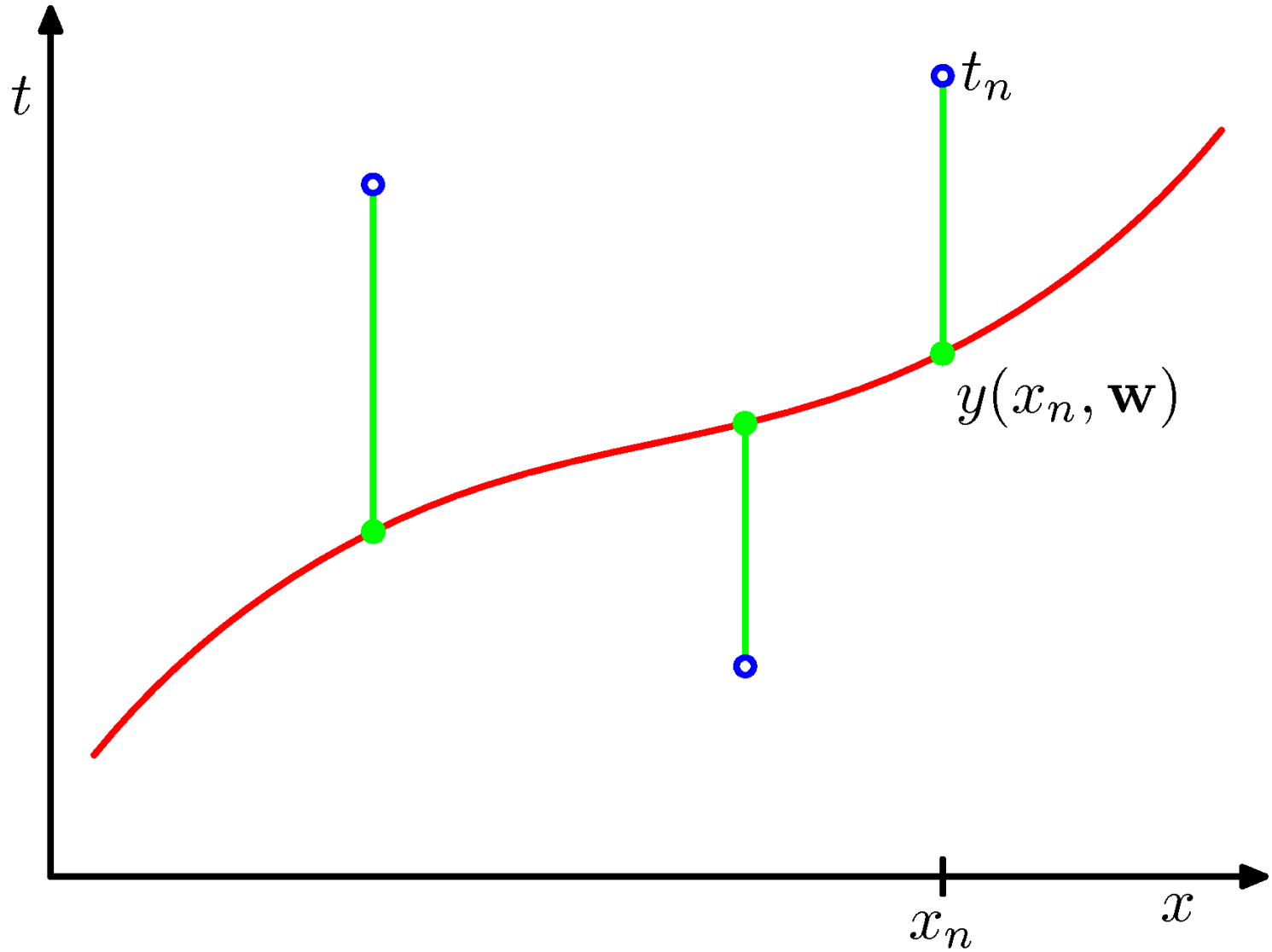
Négyzet-összeg hibafüggvény (1)

A súlyok meghatározása egy alkalmasan megválasztott célfüggvény (rizikófüggvény) minimalizálása révén történik. A leggyakrabban hasaznált rizikófüggvény az alábbi

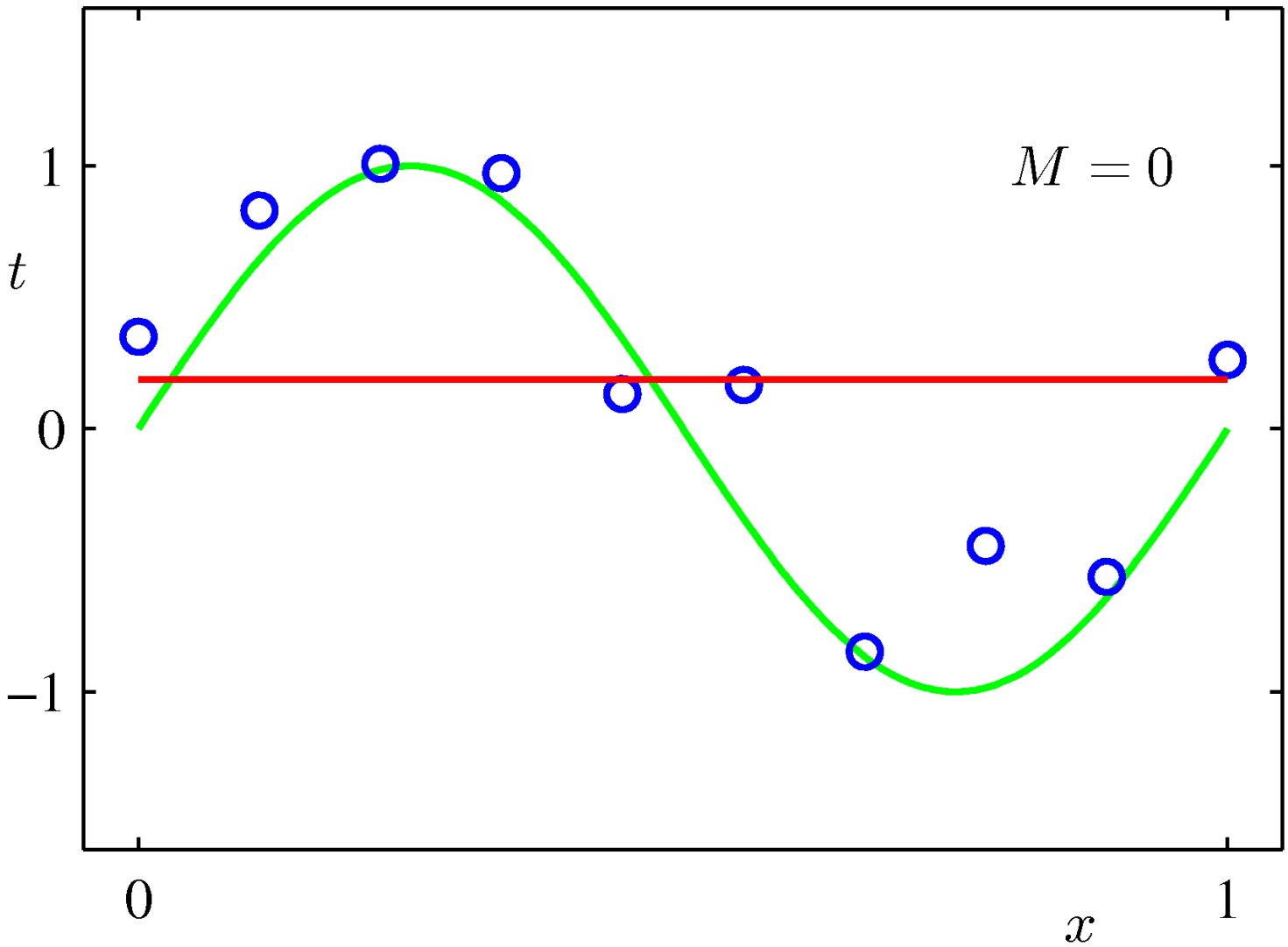
$$E(\mathbf{w}) := \frac{1}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2$$

négyzet-összeg hiba vagy négyzetes rizikó, melyet a négyzetes hibából kapunk. A négyzetes hiba az y predikció és a t célérték között az $(y - t)^2$.

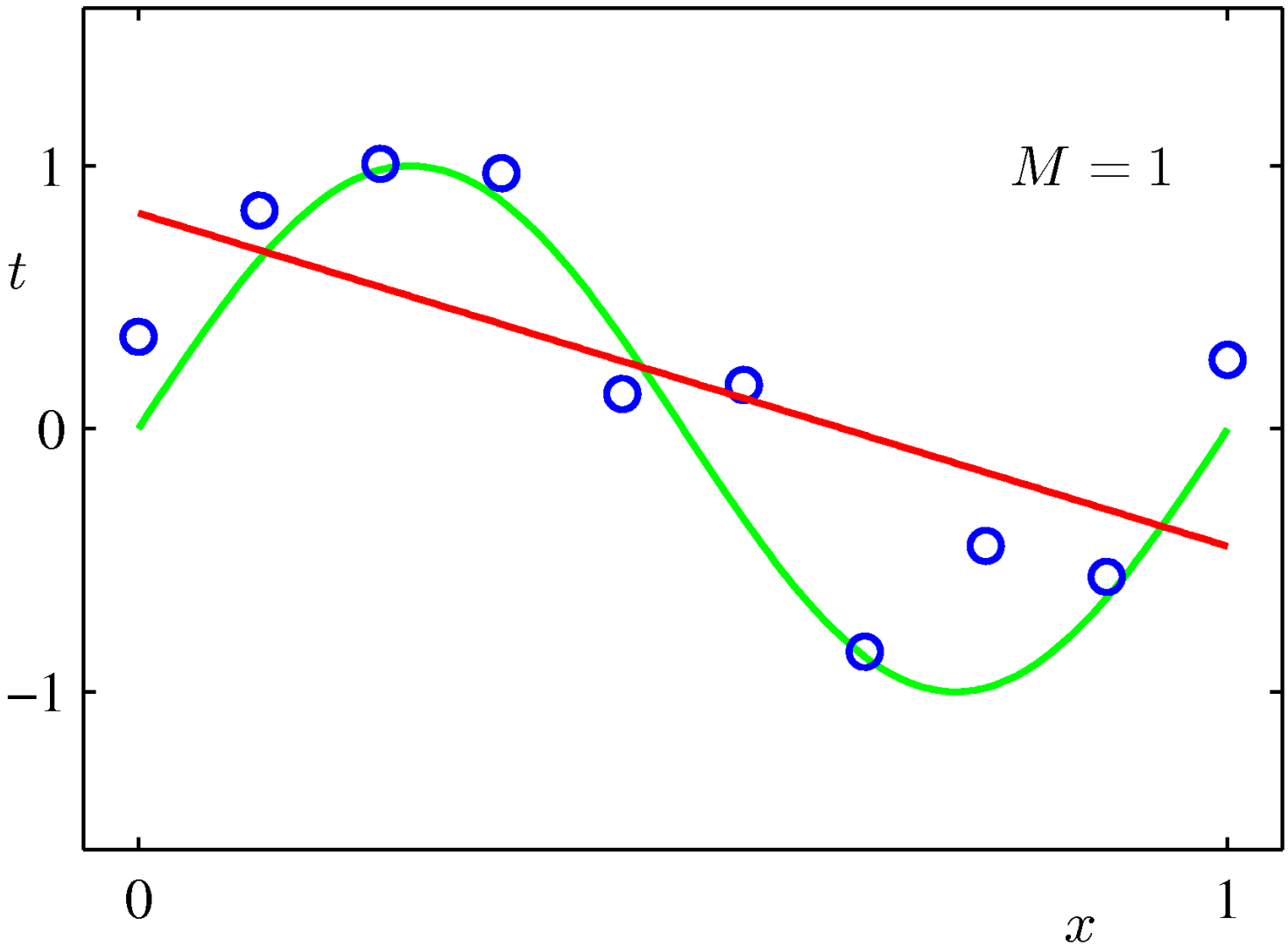
Négyzet-összeg hibafüggvény (2)



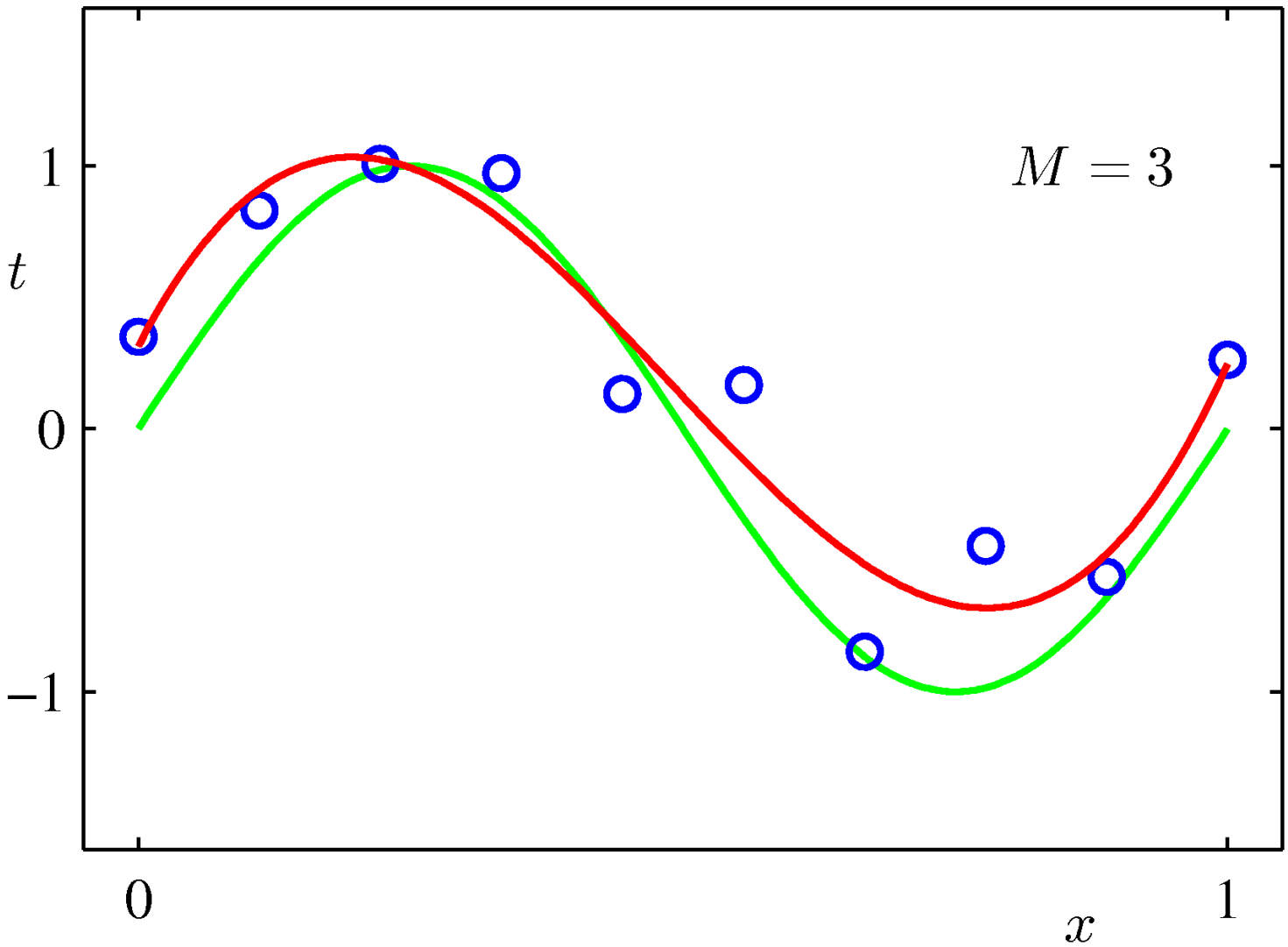
0-ad fokú polinom



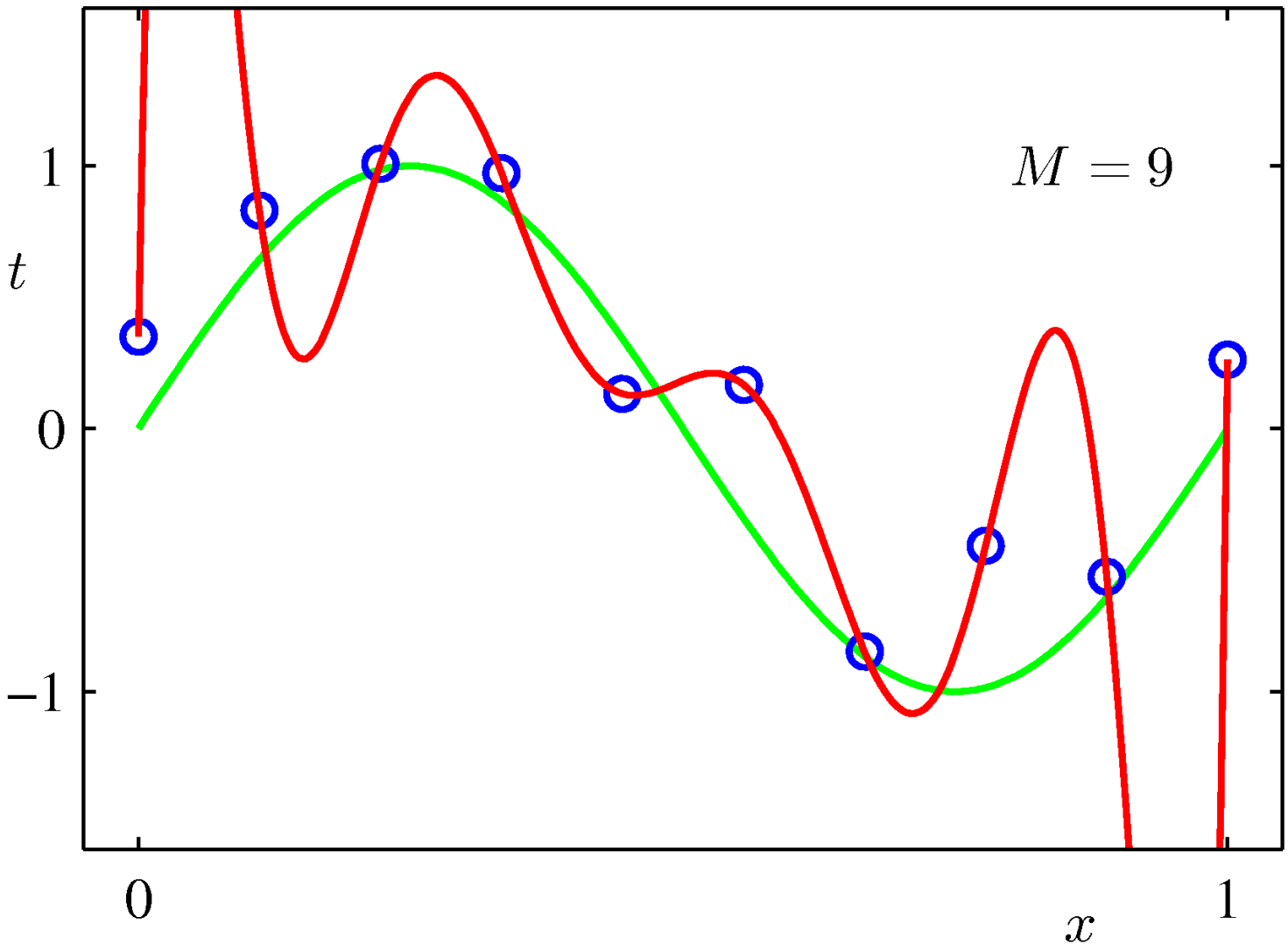
Első fokú polinom



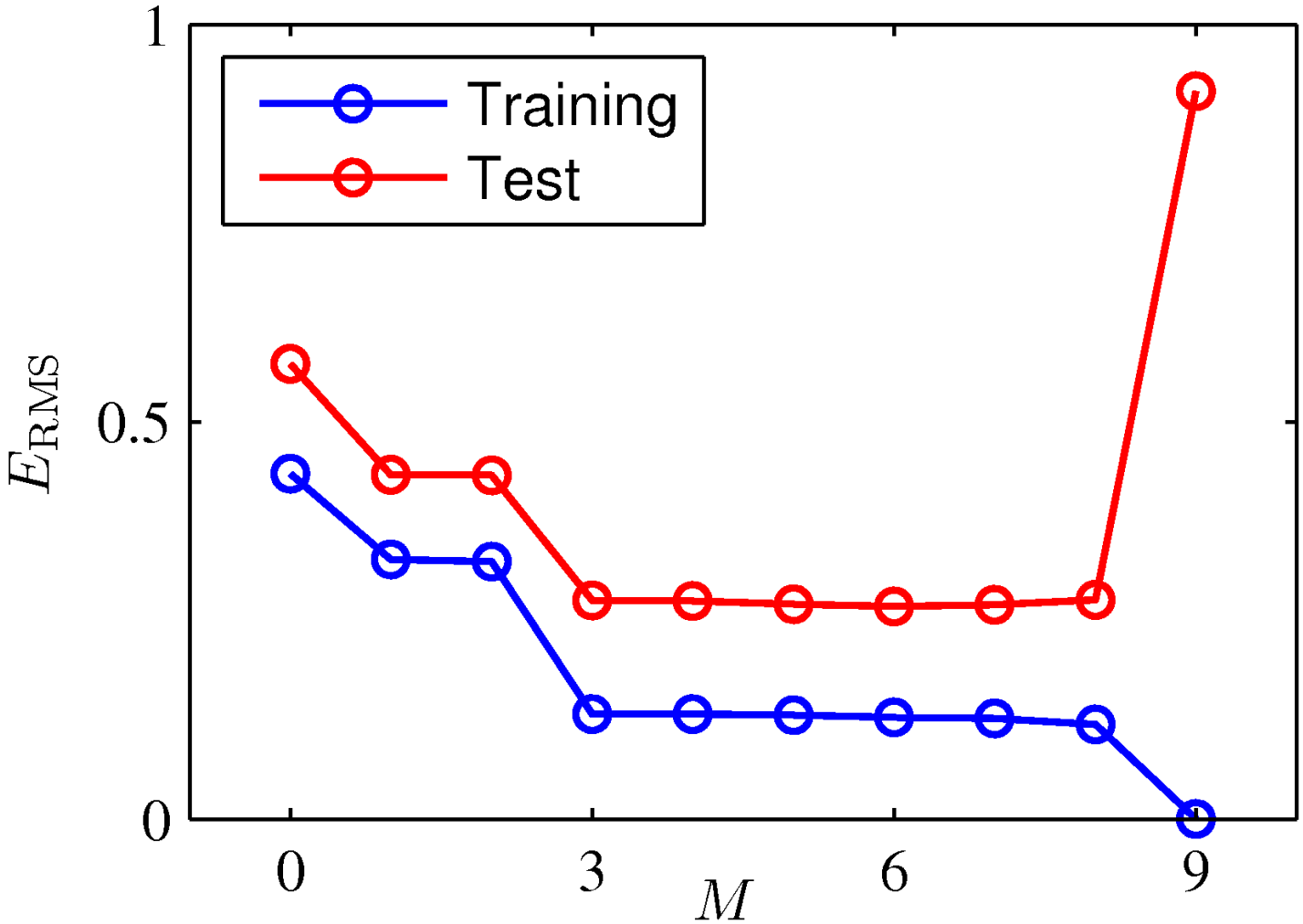
Harmad fokú polinom



9-ed fokú polinom



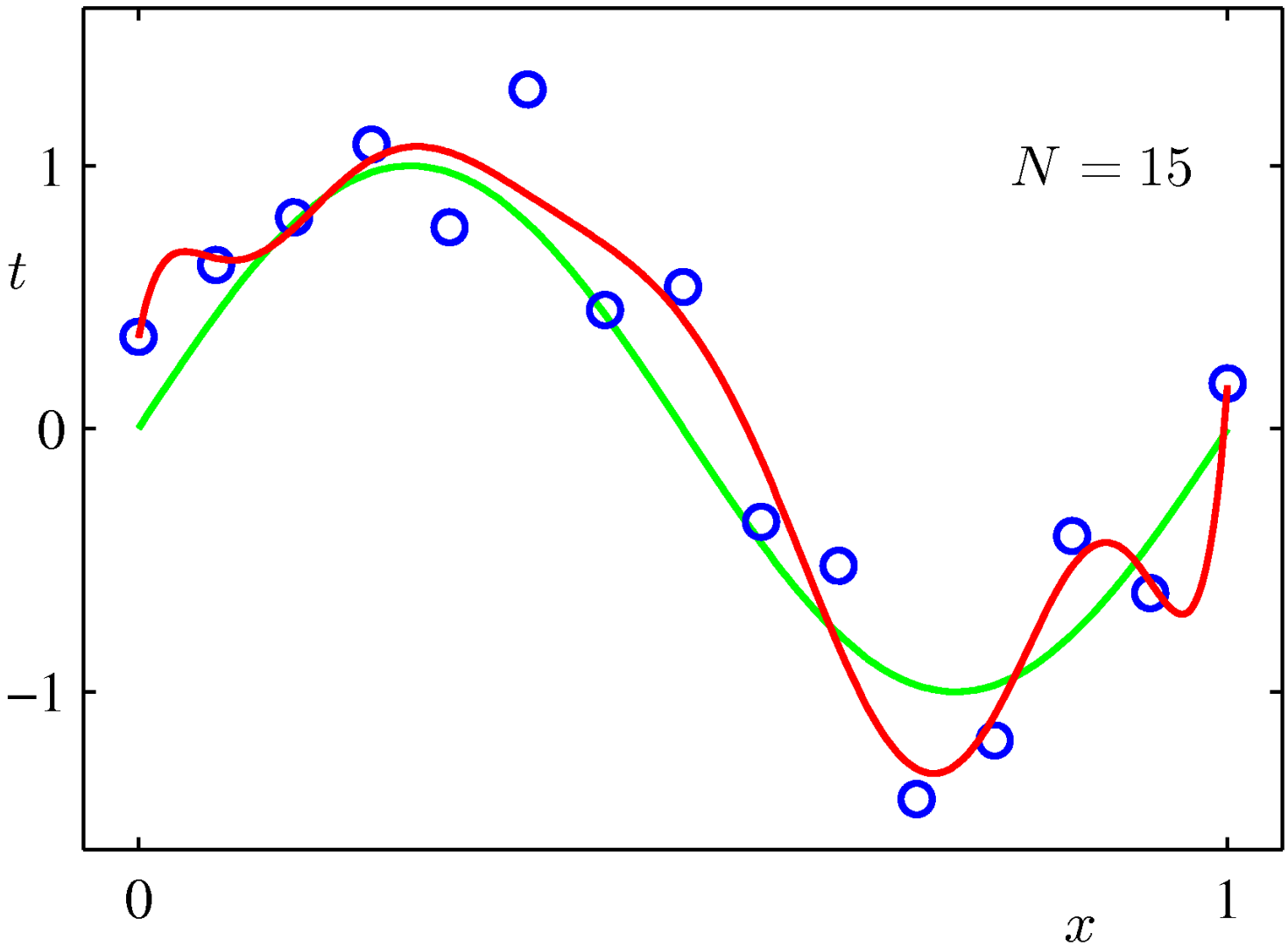
Túlillesztés



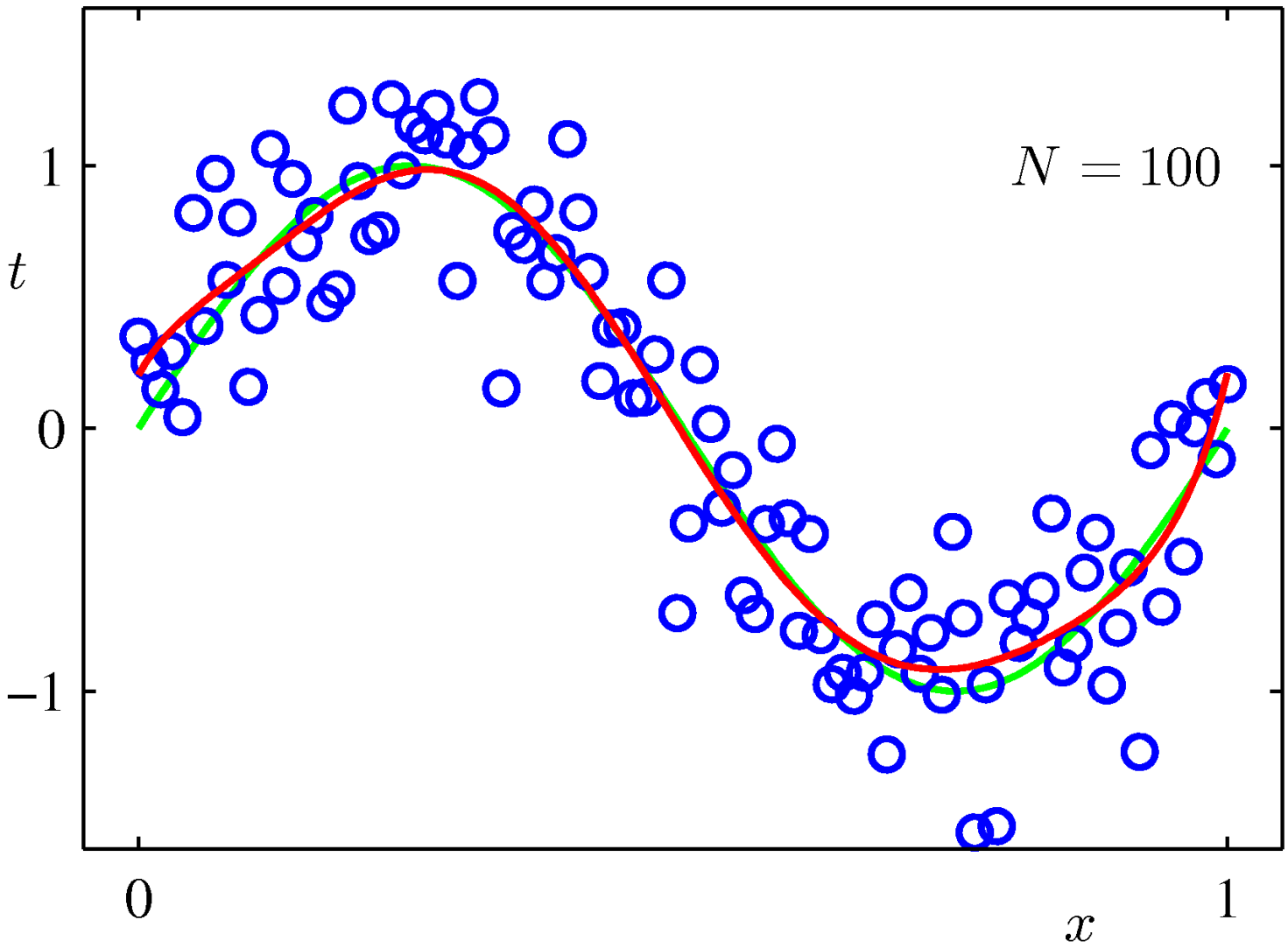
Polinom együtthatók

Súly	$M = 0$	$M = 1$	$M = 3$	$M = 9$
w_0^*	0.19	0.82	0.31	0.35
w_1^*		-1.27	7.99	232.37
w_2^*			-25.43	-5321.83
w_3^*			17.37	48568.31
w_4^*				-231639.30
w_5^*				640042.26
w_6^*				-1061800.52
w_7^*				1042400.18
w_8^*				-557682.99
w_9^*				125201.43

Adatállomány mérete ($N = 15$, $M = 9$)



Adatállomány mérete ($N = 100$, $M = 9$)



Lineáris bázis függvény modellek (1)

A polinom illesztés feladata általánosítható olyan alakban, amely már jóval flexibilisebb függvényekkel való közelítést is megenged.
Legyen

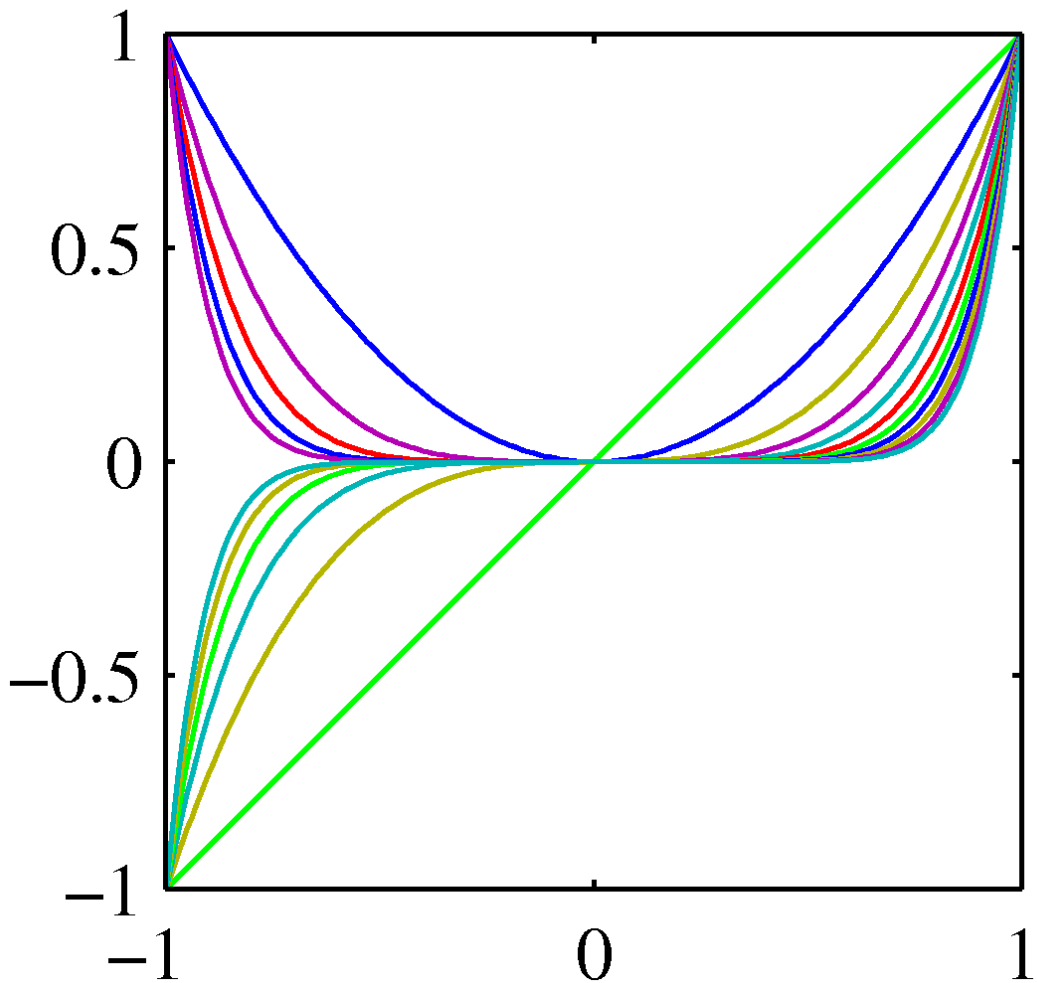
$$y(\mathbf{x}, \mathbf{w}) := \sum_{j=0}^{M-1} w_j \phi^j(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x})$$

ahol ϕ_j, $j = 0, 1, \ldots, M-1$, az ún. **bázis függvény**ek.
Rendszerint $\phi_0(\mathbf{x}) = 1$, így w_0 a torzítás.
A legegyszerűbb esetben lineáris bázisfüggvényeket használunk, ekkor $\phi_j(\mathbf{x}) = x_j$, ahol x_j az $\mathbf{x}$ vektor j-edik koordinátája.
További bázis függvények:

- polinomiális bázis függvény: $\phi_j(\mathbf{x}) = x^j$
- Gauss-féle bázis függvény: $\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$, ahol μ_j ún. hely míg s pedig skála paraméter
- Szigmoid bázis függvény: $\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right)$, ahol μ_j ún. hely míg s pedig skála paraméter, $\sigma(x) = \frac{1}{1+\exp(-a)}$ a logisztikus függvény

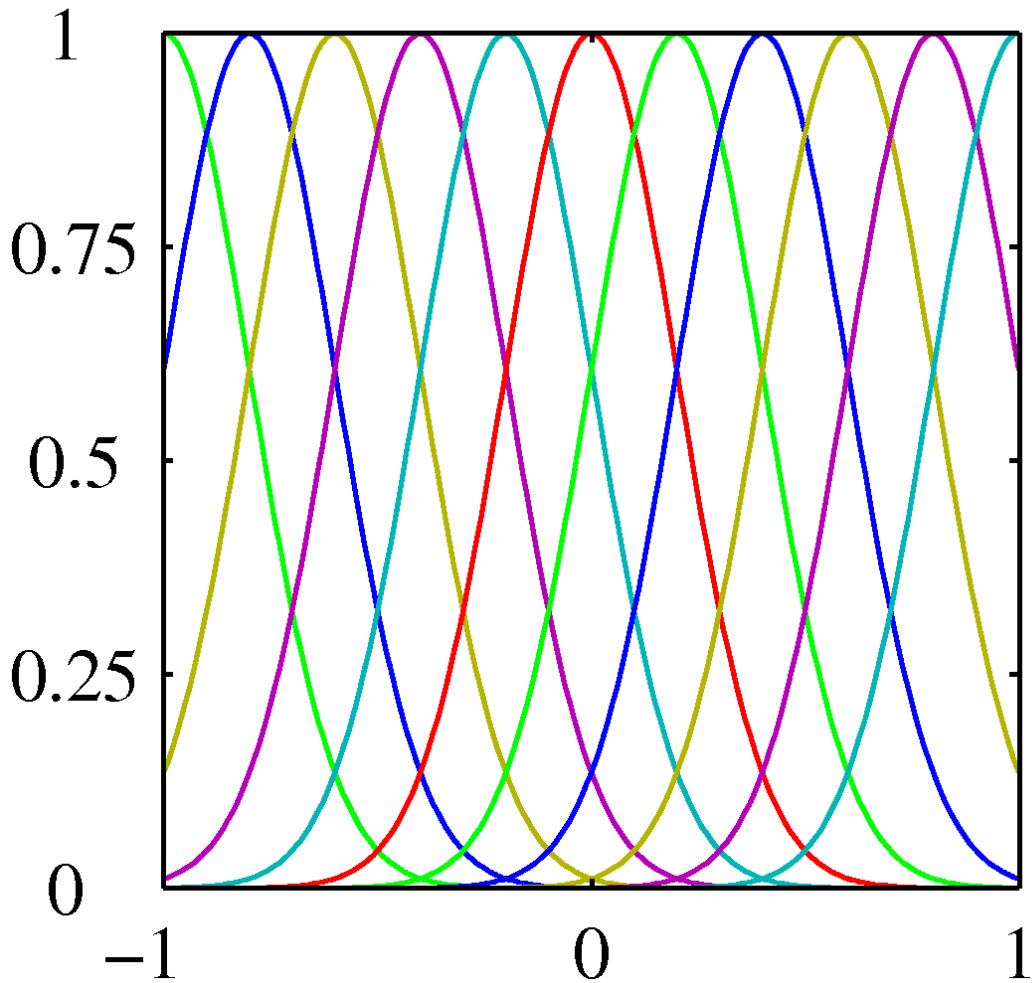
Polinom bázis függvény

Globális: kis x-beli változás kihat az összes bázis függvényre



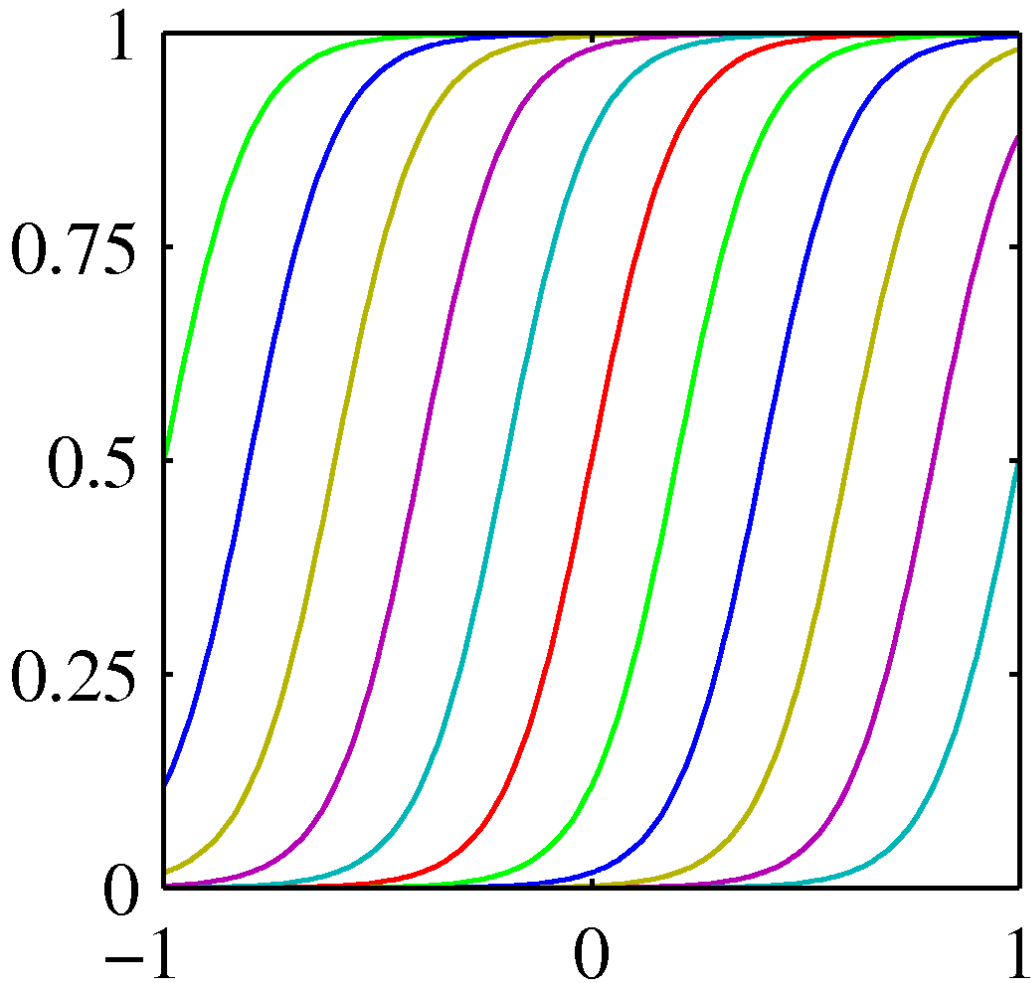
Gauss-féle bázis függvény

Lokális: kis x-beli változás csak a közeli (μ-ben) bázis föggvényekre hat ki



Szigmoid bázis függvény

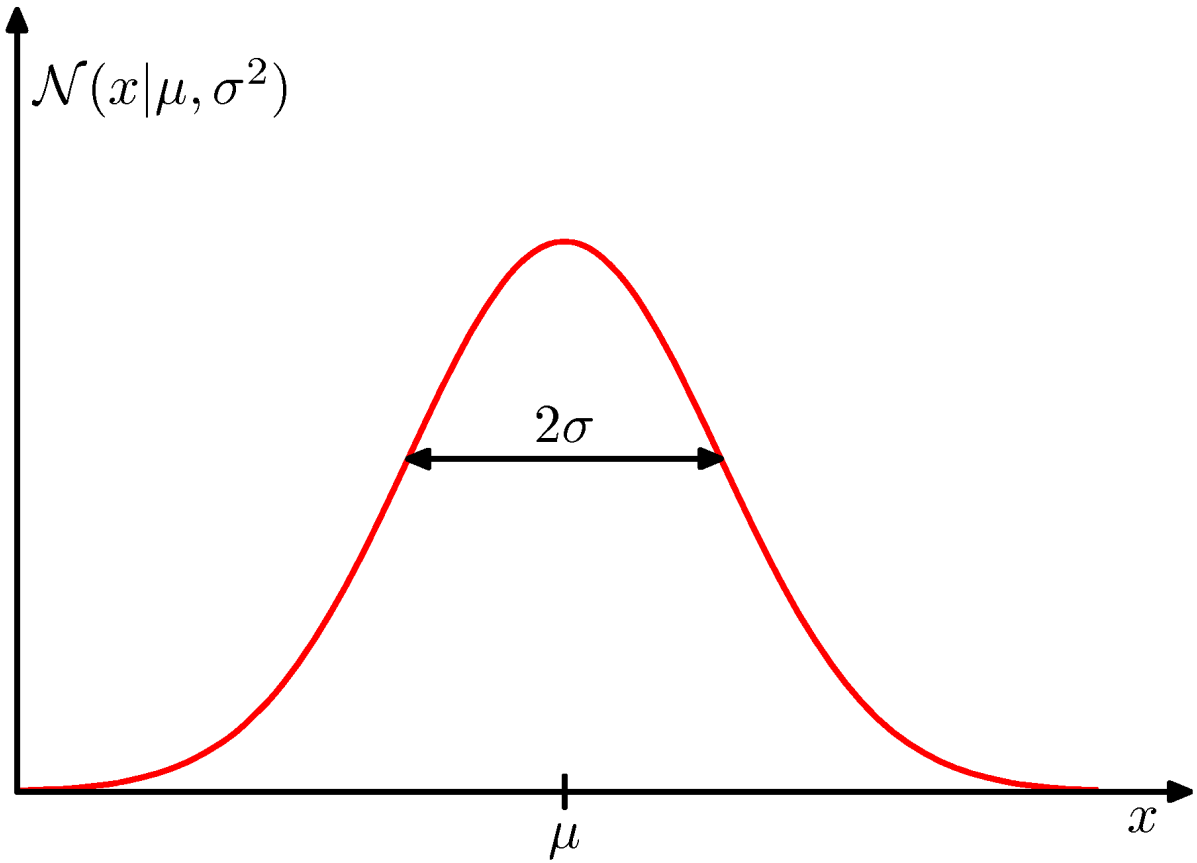
Lokális: kis x-beli változás csak a közeli (μ-ben) bázis föggvényekre hat ki



A normális eloszlás

A legszélesebb körben használt eloszlás folytonos adatok modellezésére. Sűrűségfüggvény:

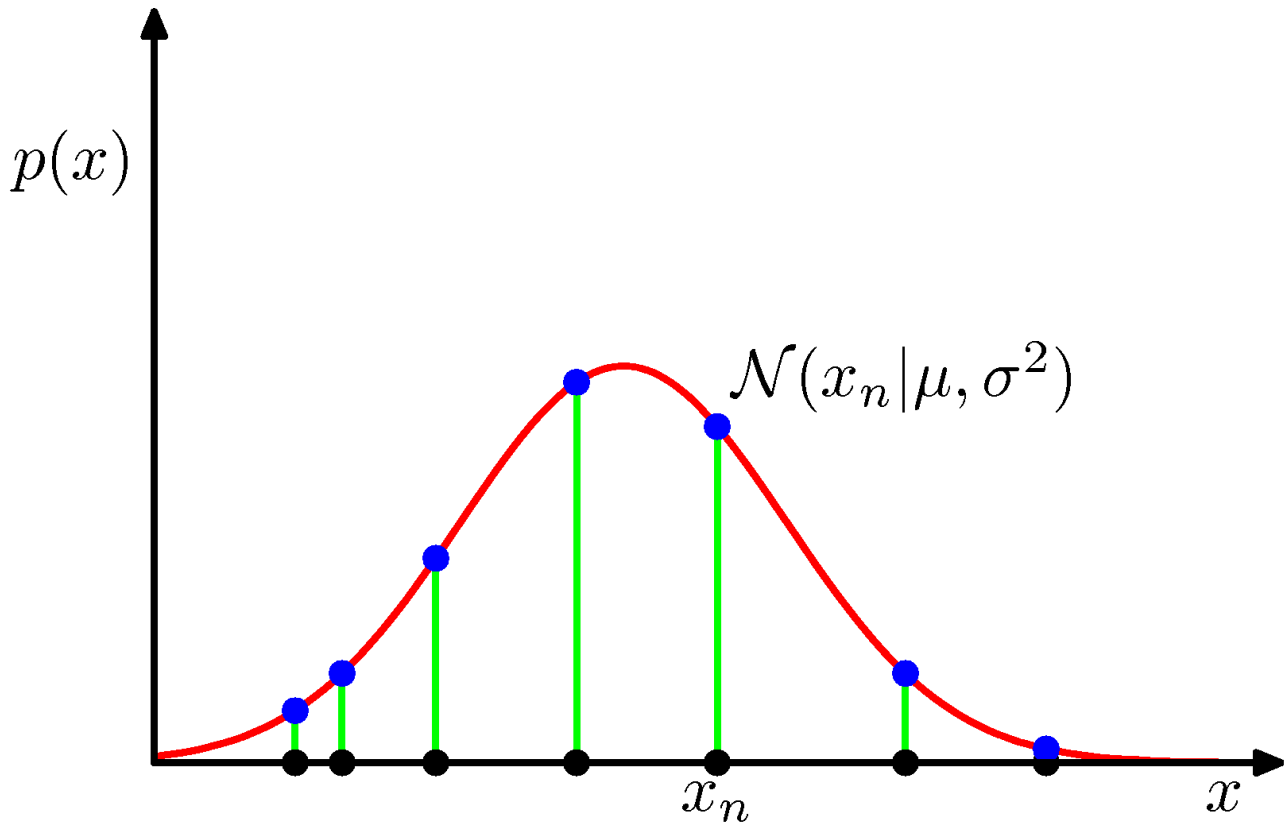
$$\mathcal{N}(x|\mu,\sigma^2) := \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



Normális eloszlás paramétereinek becslése

A maximum likelihood módszer alkalmazása. Maximalizáljuk az alábbi ún. likelihood függvényt ismert **x** megfigyelés vektor mellett a μ és σ^2 paraméterekben:

$$L(\mathbf{x}|\mu,\sigma^2) := \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2)$$



(Log)Likelihood függvény maximalizálása

$$\ln L(\mathbf{x}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

Ez μ-ben a $\sum_{n=1}^{N}(x_n-\mu)^2$ négyzetösszeg minimalizálását jelenti, ami deriválással az alábbi egyenletre vezet:

$$\sum_{n=1}^{N}(x_n-\mu) = 0$$

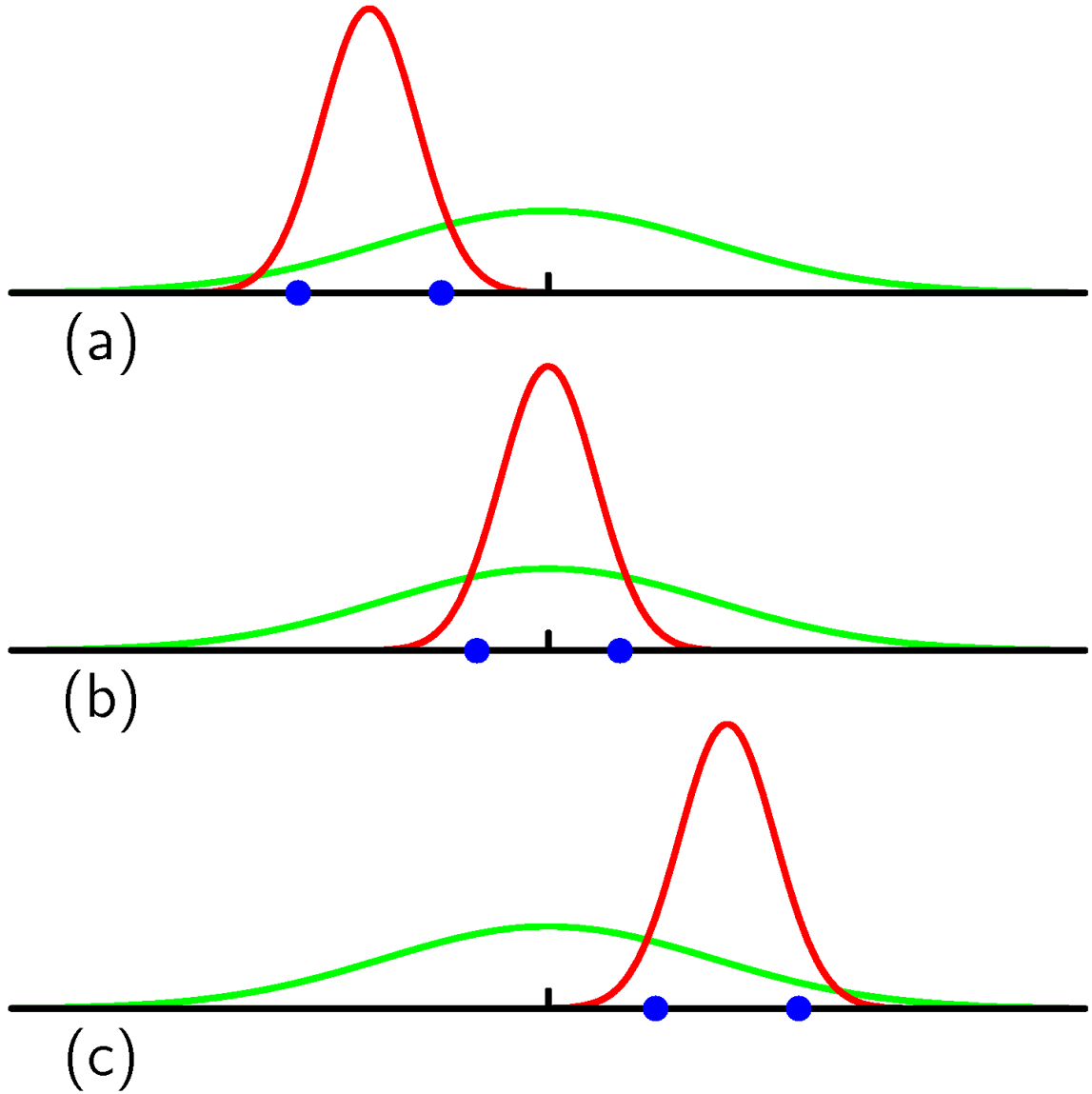
Ennek megoldása:

$$\widehat{\mu} = \frac{1}{N}\sum_{n=1}^{N}x_n$$

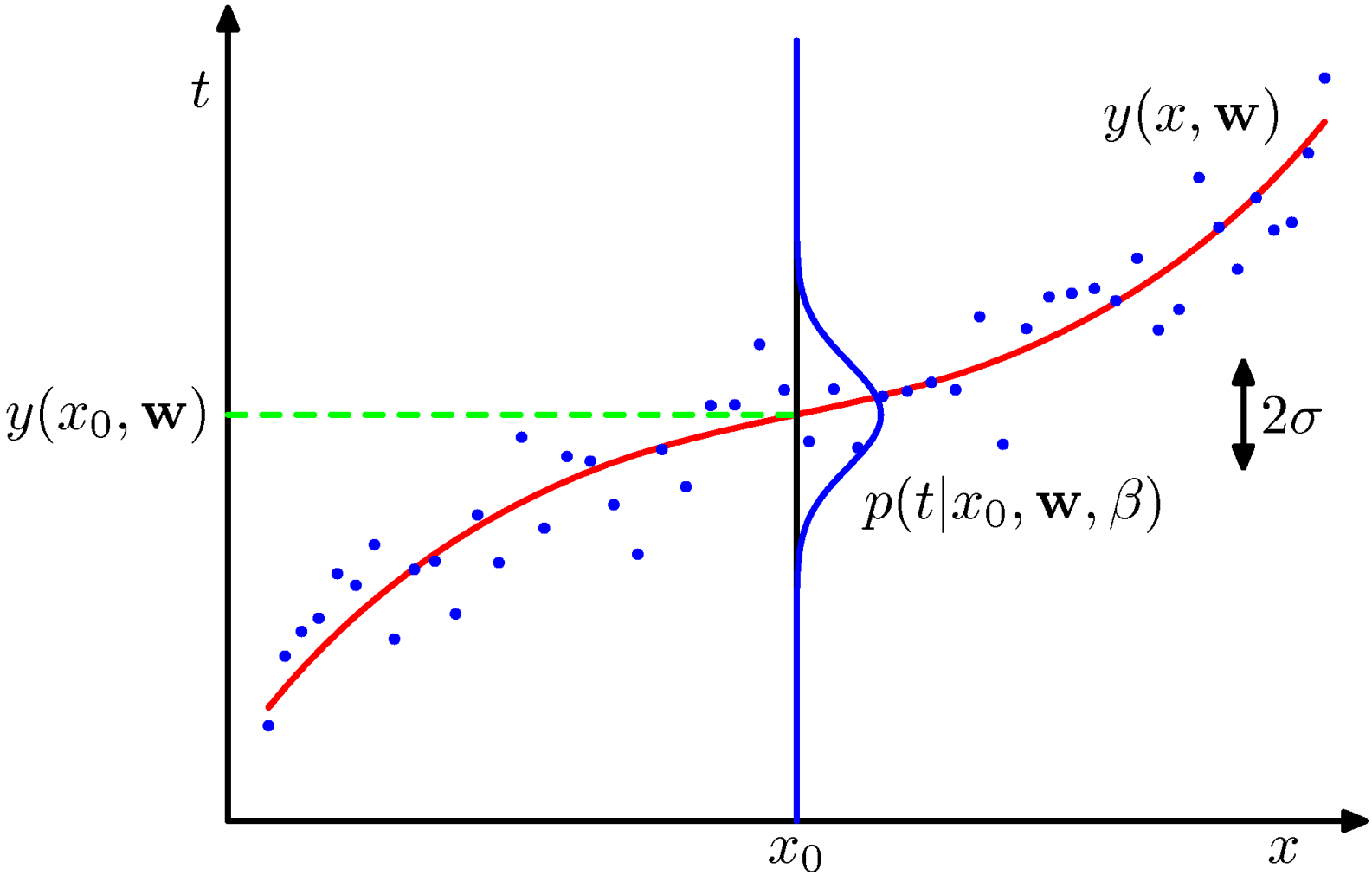
Ezután μ-be $\widehat{\mu}$-t helyettesítve és σ^2-ben maximalizálva kapjuk

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\widehat{\mu})^2$$

A becslés tulajdonságai (zöld-igazi, piros-becsült)



Sztochasztikus görbeillesztés



Maximum likelihood (1)

Legyenek adottak $(\mathbf{x}_n, t_n)$, $n = 1, 0ldots, N$, megfigyelések egy determinisztikus függvény és egy normális eloszlású hiba összegéből:

$$t_n = y(\mathbf{x}_n, \mathbf{w}) + \epsilon_n \qquad \epsilon_n \sim \mathcal{N}(0, \beta^{-1})$$

ahol β a szórásnégyzet reciproka az ún. pontosság. Ez megegyezik azzal, hogy

$$t_n \sim \mathcal{N}(y(\mathbf{x}_n, \mathbf{w}), \beta^{-1})$$

Az $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ input és $\mathbf{t} = (t_1, \ldots, t_N)^\top$ output esetén az alábbi likelihood függvényt kapjuk

$$L(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) := \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^\top \phi(\mathbf{x}_n), \beta^{-1})$$

Maximum likelihood (2)

Ismét a loglikehood függvényt tekintve az alábbit kapjuk

$$\ln L(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w})$$

ahol

$$E_D(\mathbf{w}) := \frac{1}{2} \sum_{n=1}^{N} (t_n - \mathbf{w}^\top \phi(\mathbf{x}_n))^2$$

A loglikelihood maximalizálása ekvivalens az $E_D(\mathbf{w})$ hiba négyzet-összeg minimalizálásval, amit deriválással tudunk elvégezni:

$$\nabla_{\mathbf{w}} E_D(\mathbf{w}) = \sum_{n=1}^{N} (t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)) \phi(\mathbf{x}_n)^\top = 0$$

Maximum likelihood (3)

Az ún. normálegyenletet megoldva az alábbi OLS-nek (ordinary least squares) nevezett becslés adódik a $\mathbf{w}$ paraméterre:

$$\widehat{\mathbf{w}} = (\Phi^\top \Phi)^{-1} \Phi^\top \mathbf{t} = \Phi^* \mathbf{t}$$

ahol

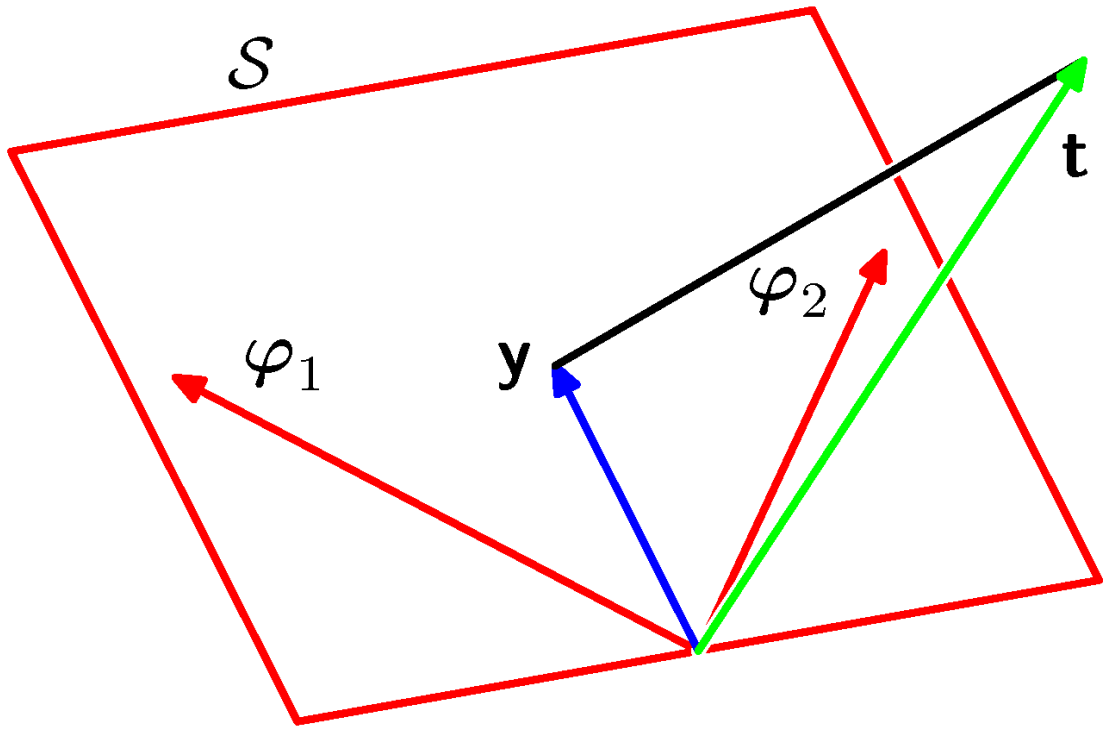
$$\Phi := \begin{bmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{bmatrix}$$

egy $N \times M$-es mátrix. Φ^* az ún. Moore-Penrose pszeudo-inverz. Ha a β paraméterre is maximalizálunk, akkor kapjuk, hogy

$$\beta^{-1} = \frac{1}{N} \sum_{n=1}^{N} (t_n - \widehat{\mathbf{w}}^\top \boldsymbol{\phi}(\mathbf{x}_n))$$

A legkisebb négyzetek geometriája

Legyen $\mathbf{y} = \Phi\widehat{\mathbf{w}} = [\phi_0, \ldots, \phi_{M-1}]\widehat{\mathbf{w}}$
$\mathbf{y} \in \mathcal{S} \subseteq \mathcal{T}$ $\quad$ $\mathbf{t} \in \mathcal{T}$
$\mathcal{S}$ M-dimensional, $\mathcal{T}$ N-dimensional ($M \leq N$)
Az $\mathcal{S}$ síkot a $\phi_0, \ldots, \phi_{M-1}$ vektorok (Φ oszlopai) feszítik fel



Regularizált legkisebb négyzetek (1)

Tekintsük az alábbi hiba (rizikó) függvényt:

$$E_D(\mathbf{w}) + \lambda E_{\mathbf{w}}(\mathbf{w}),$$

ahol λ a regularizációs együttható. Négyzetes hibafüggvénnyel és kvadratikus regulátorral kapjuk, hogy az alábbi kifejezést kell minimalizálni:

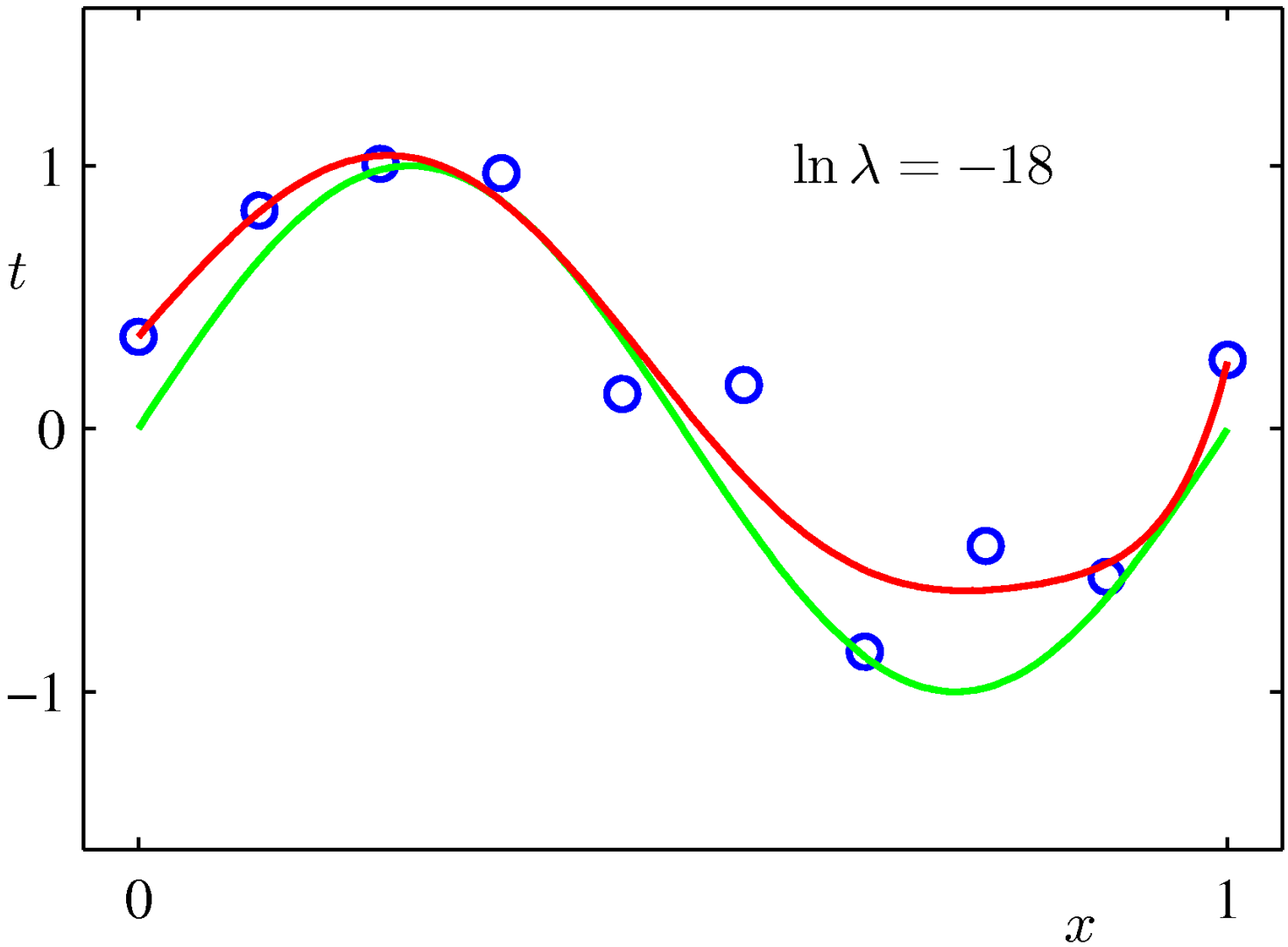
$$\frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\top \phi(\mathbf{x}_n))^2 + \frac{\lambda}{2}\mathbf{w}^\top \mathbf{w}$$

A megoldás:

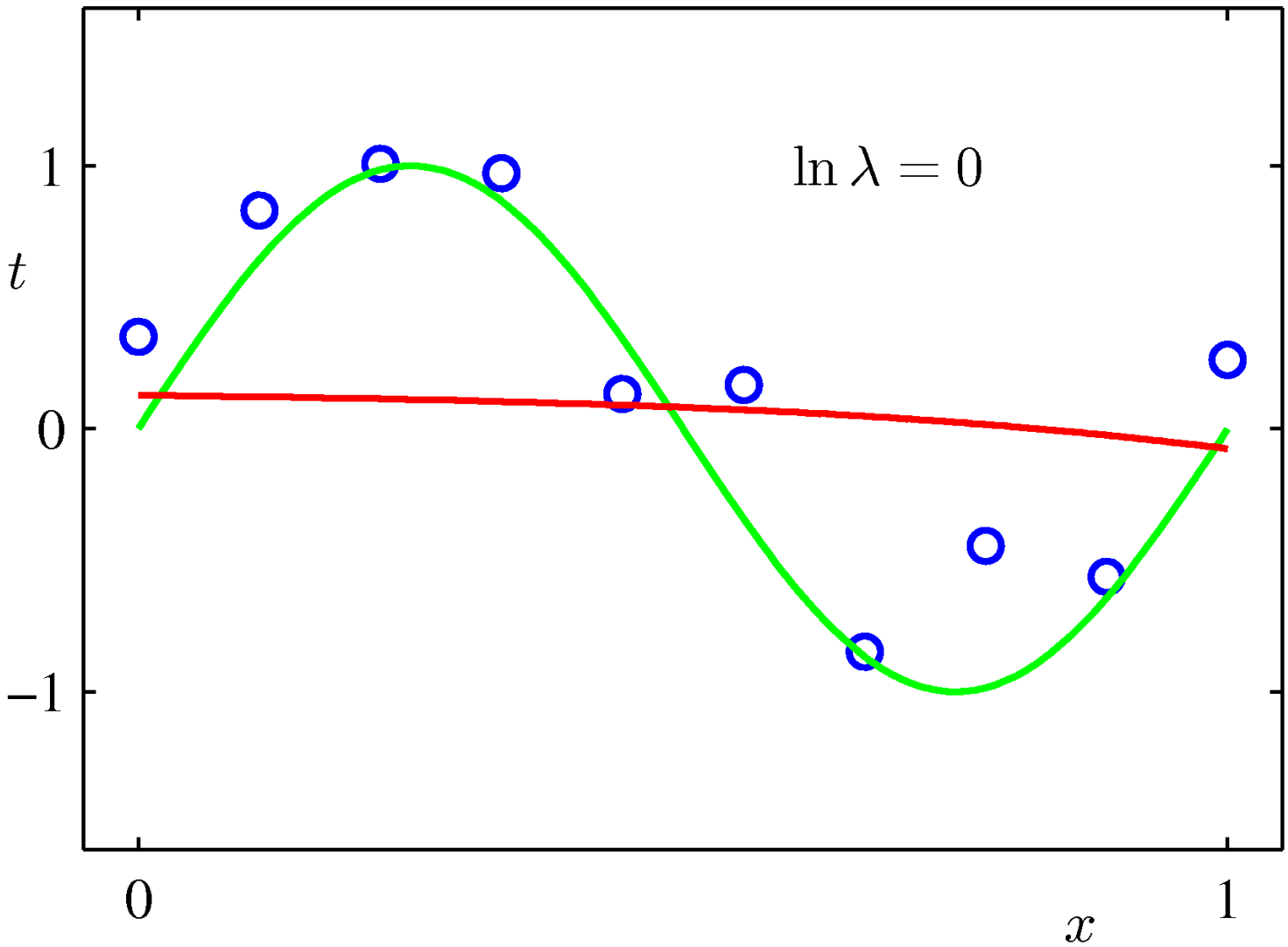
$$\widehat{\mathbf{w}}_\lambda = (\lambda I + \Phi^\top \Phi)^{-1}\Phi^\top \mathbf{t}$$

Ridge regresszió, segít ha az OLS-beli inverz nem létezik.

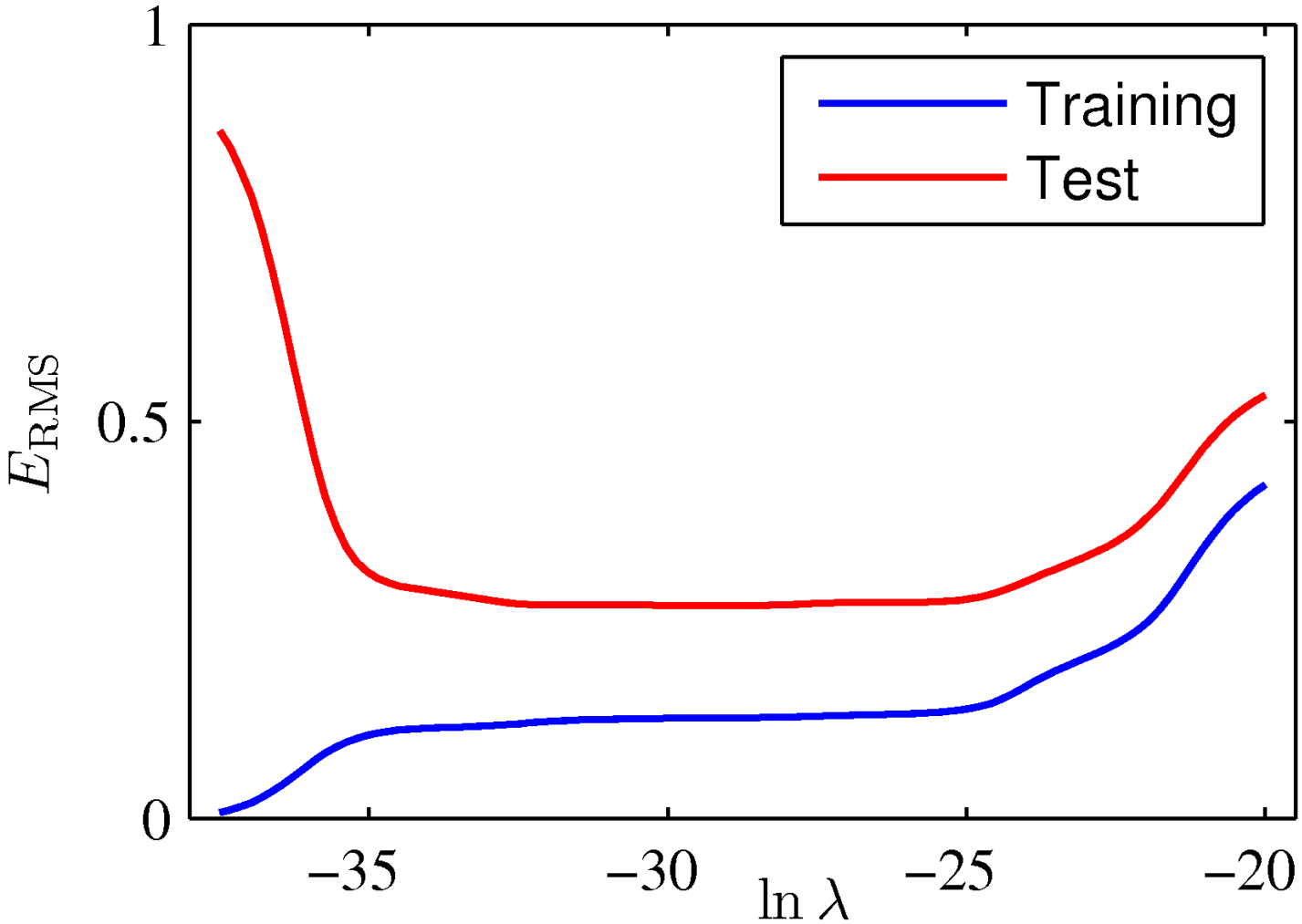
Regularizáció



Regularizáció



Regularizáció



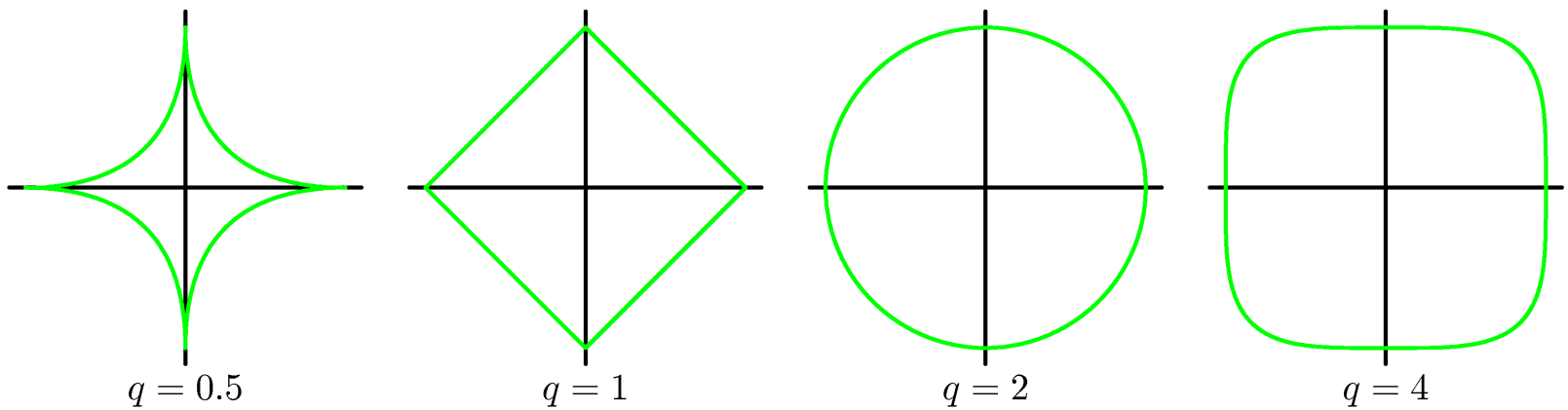
Polinom együtthatók

Súly	$ln\lambda = -\infty$	$ln\lambda = -18$	$ln\lambda = 0$
w_0^*	0.35	0.35	0.13
w_1^*	232.37	4.74	-0.05
w_2^*	-5321.83	-0.77	-0.06
w_3^*	48568.31	-31.97	-0.05
w_4^*	-231639.30	-3.89	-0.03
w_5^*	640042.26	55.28	-0.02
w_6^*	-1061800.52	41.32	-0.01
w_7^*	1042400.18	-45.95	-0.00
w_8^*	-557682.99	-91.53	0.00
w_9^*	125201.43	72.68	0.01

Regularizált legkisebb négyzetek (2)

Általánosabb regularizálót is használhatunk ($q > 0$):

$$\frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\top \phi(\mathbf{x}_n))^2 + \frac{\lambda}{2}\sum_{j=0}^{M-1}|w_j|^q$$



Elnevezés: $q = 1$ lasso, $q = 2$ kvadratikus

Regularizált legkisebb négyzetek (3)

A lasso hajlamosabb ritkább (sparse) megoldásokat előállítani mint a kvadratikus regularizáló.