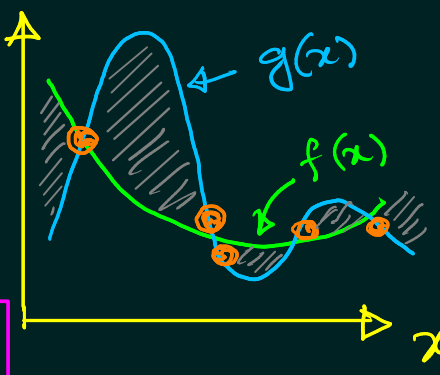
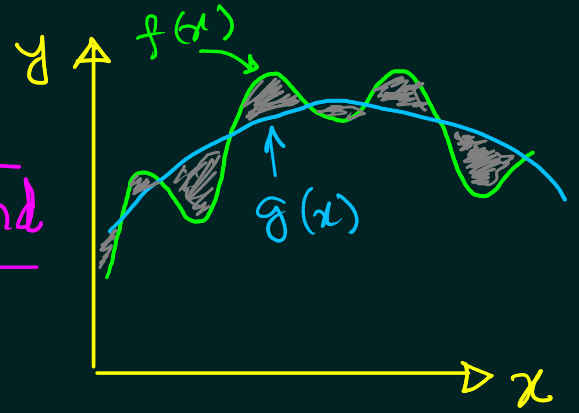
Overfitting:

1. Stochastic Noise
2. Deterministic Noise



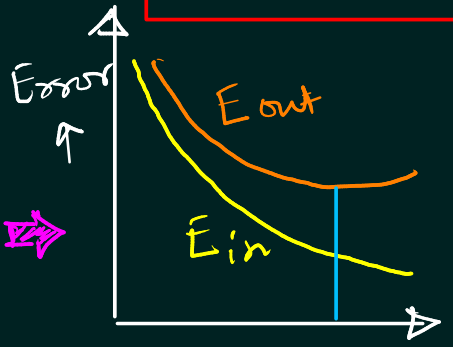
Data with Noise fitting ①

and

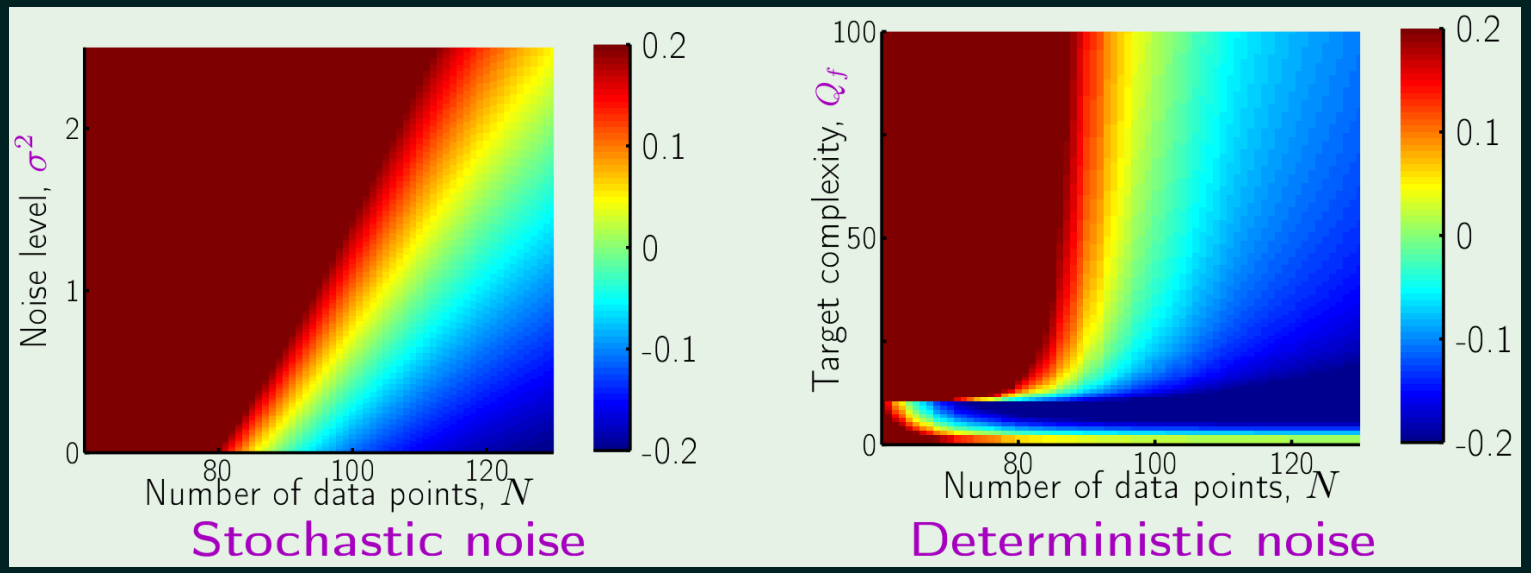


Higher order Data fitting ②

SUMMARY



Overfitting happens



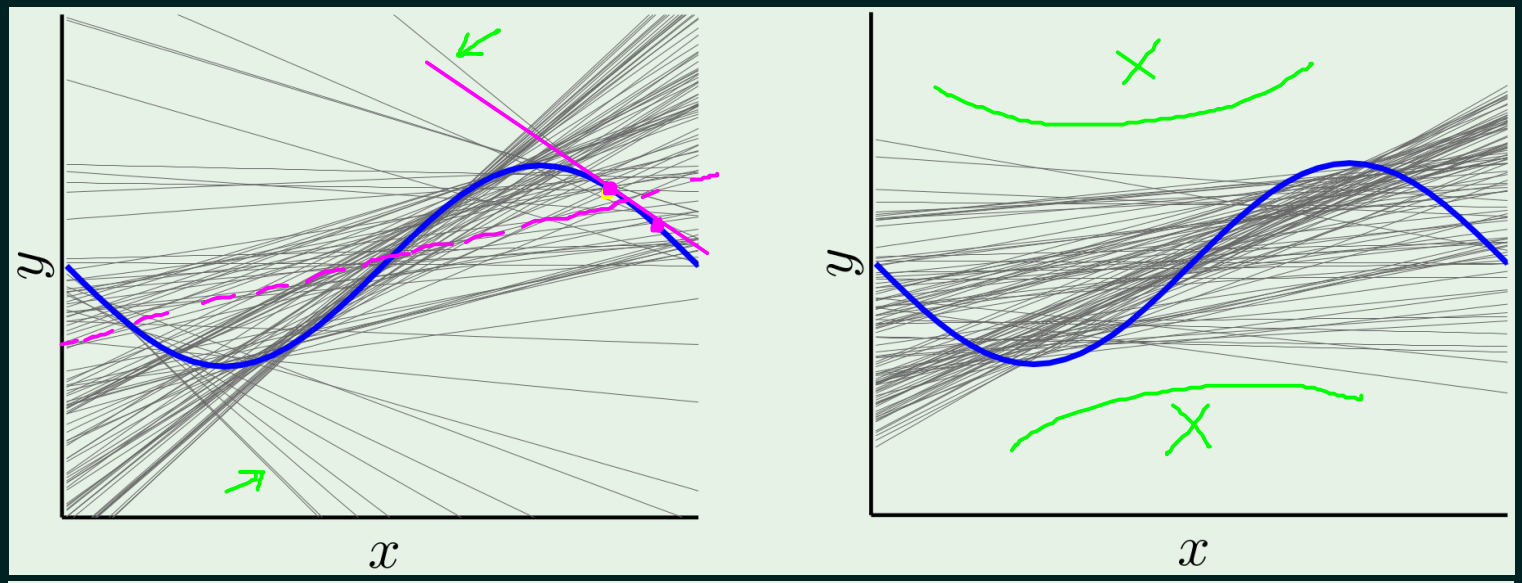
- ↑ Training Data → ↓ Overfitting
- ↑ Noise level → ↑ overfitting
- ↑ Target Complexity → ↑ Overfitting

→ Relation with VC-Dimension:
↳ Generalization Bound ⇒ Allows it ✓ but Never Predicts this ??

→ Overfitting Remedy:
① REGULARIZATION + ② VALIDATION

THIS WEEK'S COVERAGE

Regularization to Rescue



without regularization

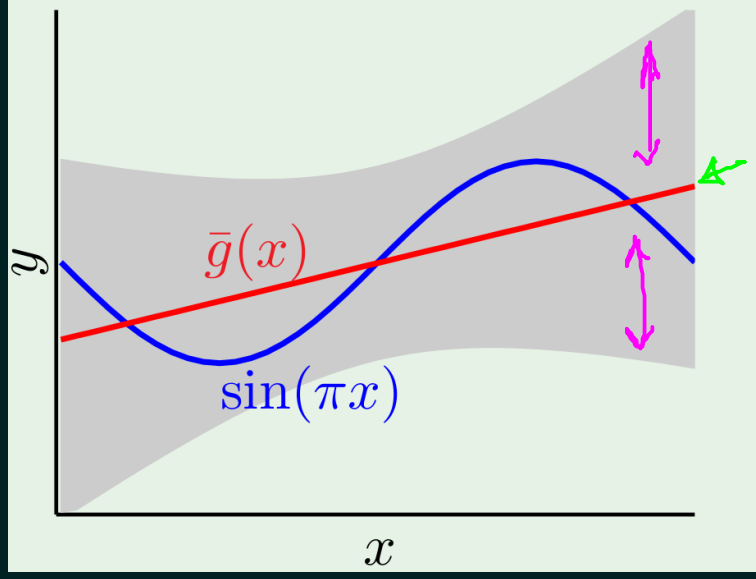
with regularization

Not Allow all slopes

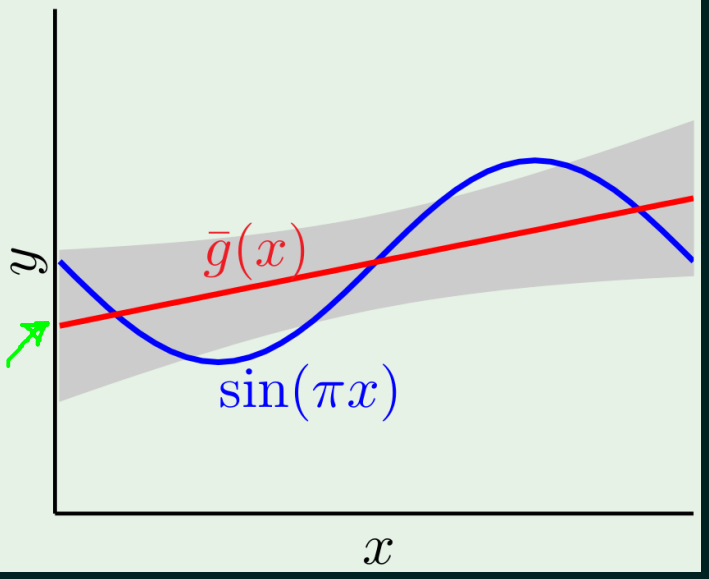
$w f(x) = ax + b$

$\checkmark$

$f(x) = c$



bias = 0.21 var = 1.69



bias = 0.23 var = 0.33

Constraining Solution

$\mathcal{H}_Q = \{ \sum_{q=0}^{Q} w_q L_q(x) \}$ $w_q \to 0, 1$

$(x_1, y_1) \quad \cdots \quad (x_N, y_N)$

$\Downarrow$

$(z_1, y_1) \quad \cdots \quad (z_N, y_N)$

$$Z = \begin{bmatrix} 1 \\ L_1(x) \\ \vdots \\ L_Q(x) \end{bmatrix}$$

$$\text{Min. } E_{in}(w) = \frac{1}{N} \sum_{n=1}^{N} (w^T z_n - y_n)^2$$

$$= \frac{1}{N} (Zw - y)^T (Zw - y)$$

$$w_{lin} = (Z^T Z)^{-1} Z^T y$$

linear Regression

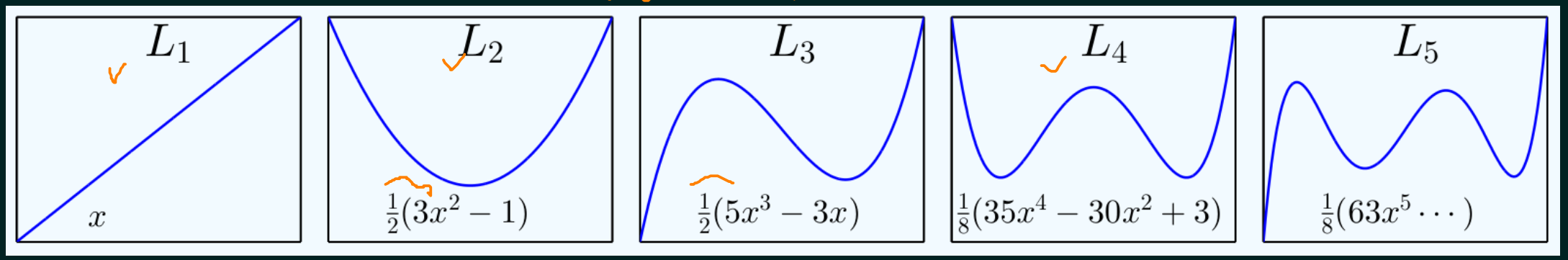
$\mathcal{H}_2$

$\mathcal{H}_{10}$

$\mathcal{H}_q$ with $w_q = 0$ for $q > 2$ (Hard)

Legendre Polynomials:

[Unconstrained Fitting of N points]



$\mathcal{H}_Q = \left\{ \sum_{q=0}^{Q} w_q L_q(x) \right\}$ soft constraint : $\sum_{q=0}^{Q} w_q^2 \leq C$ $\mathcal{H}_2$

$w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$

Minimize $E_{in}(w) = \frac{1}{N}(Zw - Y)^T (Zw - Y)$

Subject to $w^T w \leq C$ → w_{reg}

Reg

$\nabla E_{in}(w_{reg}) \propto -w_{reg}$

Real

$= -2\frac{\lambda}{N} w_{reg}$

$\Rightarrow \nabla E_{in}(w_{reg}) + 2\frac{\lambda}{N} w_{reg} = 0$

Min. $E_{in}(w) + \frac{\lambda}{N} w^T w$

unconst

s.t. $\lambda \geq 0$

$\lambda > 0$

$\lambda = 0$

$w_{reg} = 0$

KKT

$C \downarrow$ $C \uparrow$

$\lambda \uparrow$ $\lambda \downarrow$

Minimize $E_{aug}(w) = E_{in}(w) + \frac{\lambda}{N} w^T w$

$= \frac{1}{N}\left[(Zw - y)^T (Zw - y) + \lambda w^T w \right]$

Q.P Tech.

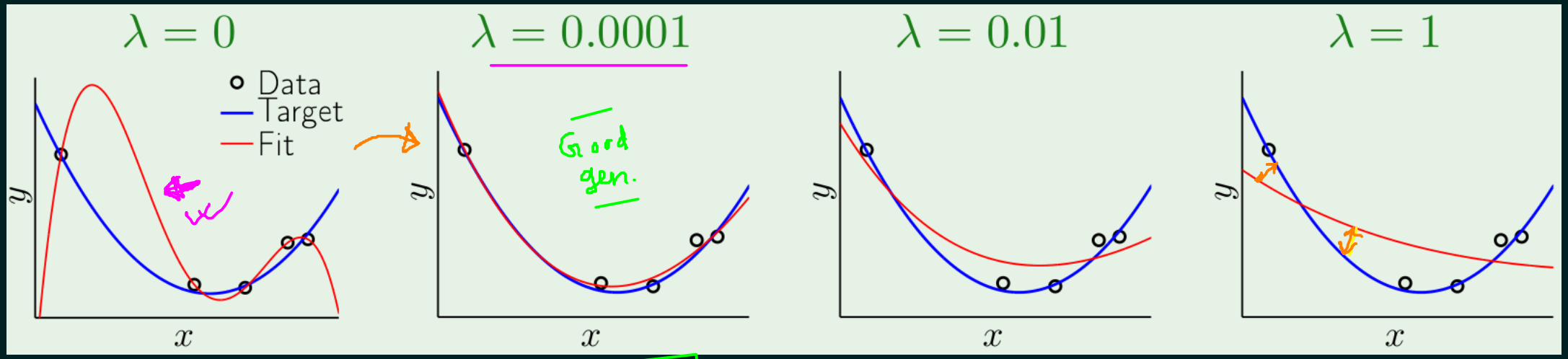
$\nabla E_{in}(w) = 0 \Rightarrow Z^T (Zw - y) + \lambda w = 0$

(Reg) $\Rightarrow w_{reg} = (Z^T Z + \lambda I)^{-1} Z^T y$

Opposed to; $w_{lin} = (Z^T Z)^{-1} Z^T y$ (W/o Reg)

$\lambda \uparrow \inf \Rightarrow w_{reg} \approx 0$ / $\lambda \downarrow \approx 0 \Rightarrow w_{reg} \approx w_{lin}$

→ More $\lambda \Rightarrow$ flatter curves (Smoother)



overfitting → Reg 1 → Reg 1' → ⋯ → underfitting

▷ Weight Decay :

Gradient Descent

$$w(t+1) = w(t) - \eta \left[\nabla E_{in}(w(t)) + \frac{2\lambda}{N} w(t) \right]$$

$$= w(t) \left[1 - 2\eta \frac{\lambda}{N} \right] - \eta \nabla E_{in}(w(t))$$

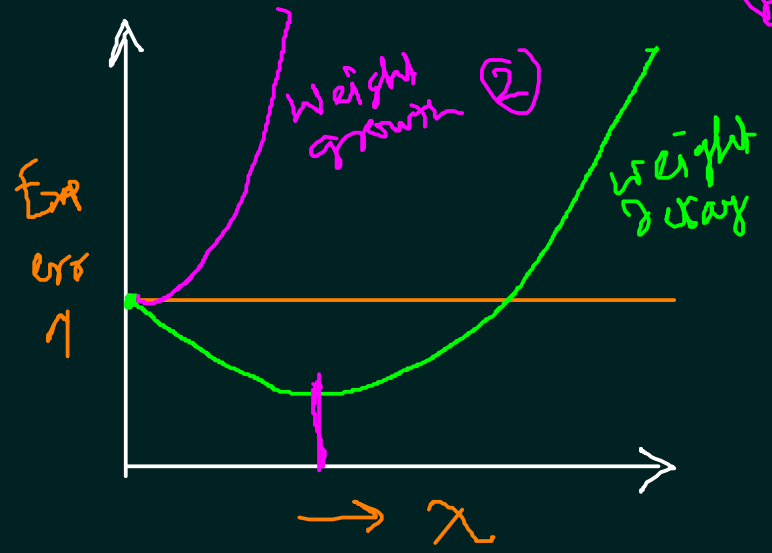
decay

$$\left[w^T w = \sum_{l=1}^{L} \sum_{i=0}^{d^{l-1}} \sum_{j=1}^{d^l} \left(w_{ij}^l \right)^2 \right]$$

$$\sum_{q=0}^{Q} \gamma_q w_q^2 \leq C$$

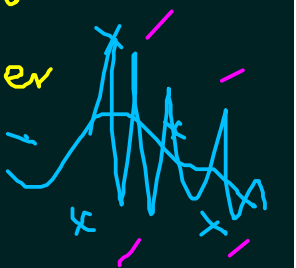
$\gamma_q = 2^q$ → low order poly

$\gamma_q = 2^{-q}$ → high order poly



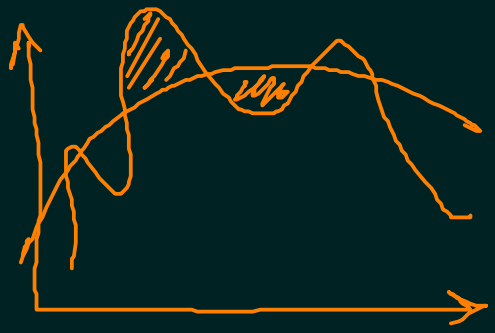
① $\sum w_q^2 \leq C$ → const weight smaller

② $\geq C$ ✕

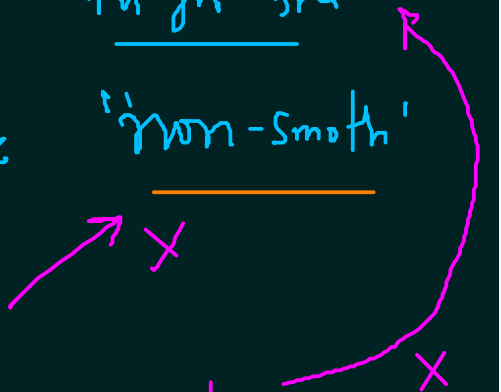


→ Stochastic Noise "high-fru"

→ Deterministic Noise "non-smoth"



Practical Rule ⇒ "Smoother hypothesis"



$H_2 > H_{10}$

[OCCAM RAZOR]

$$E_{aug}(h) = E_{in}(h) + \frac{\lambda}{N}\Omega(h)$$

$$E_{out}(h) \leq E_{in}(h) + \Omega(\mathcal{H})$$

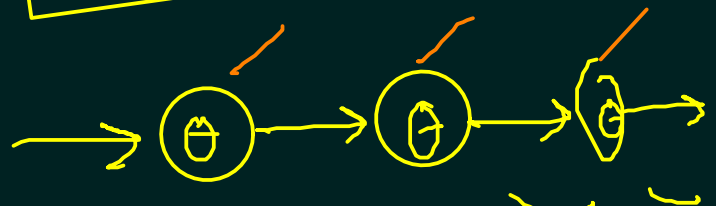
$E_{aug}(h)$ is closer est. / better than E_{in} as proxy for E_{out}

linear

↓

logical

$w_i x_i$



$$\Omega(w) = \sum \frac{(w_{ij}^{(l)})^2}{\beta^2 + (w_{ij}^{(l)})^2}$$

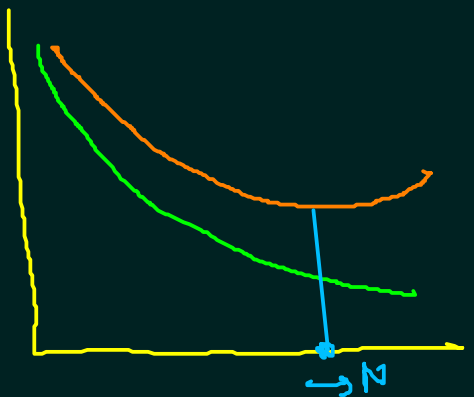
Reg → λ (validation)

→ Z (Legendre poly)

soft weight elimination

validation

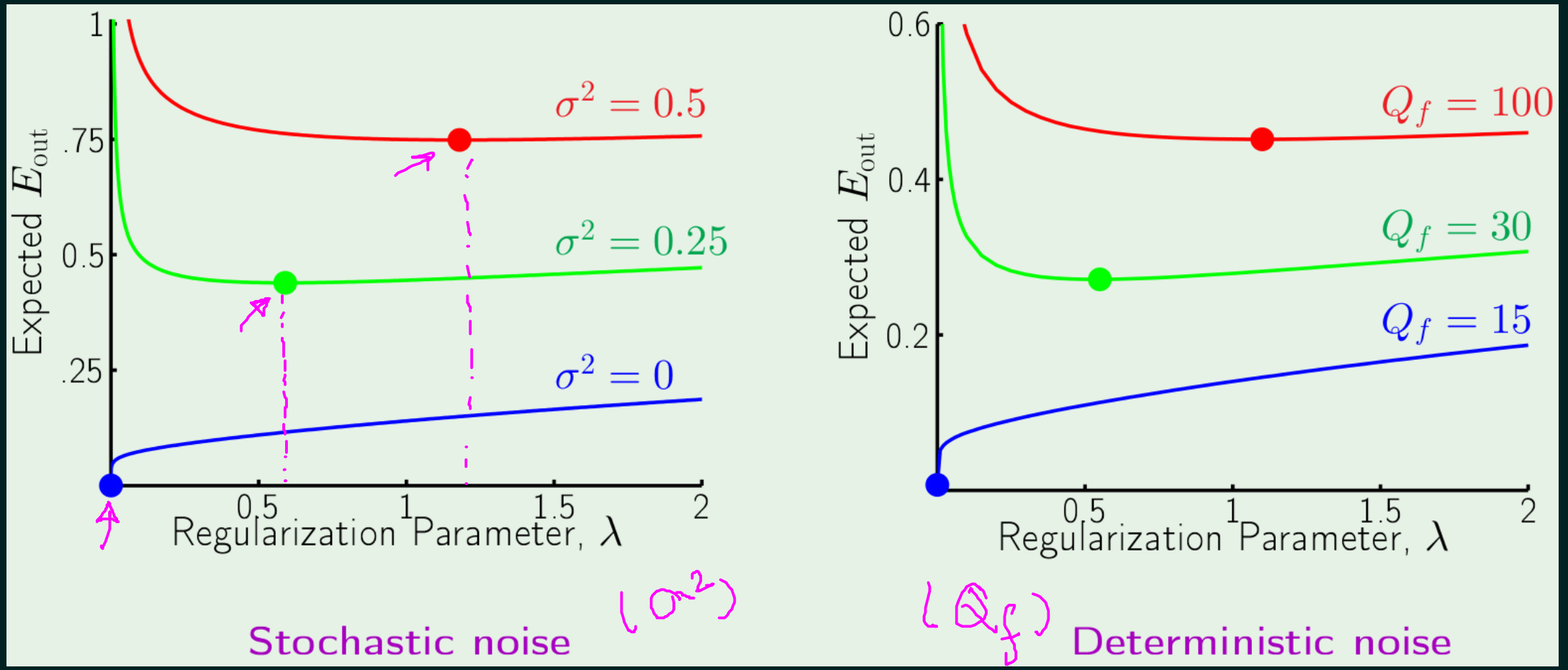
$$\mathcal{D}_{train} = \mathcal{D}_t \cup \mathcal{D}_v$$



Optimal λ

→ Deterministic
→ Stochastic

Validation λ



(σ^2)

Stochastic noise

(Q_f)

Deterministic noise