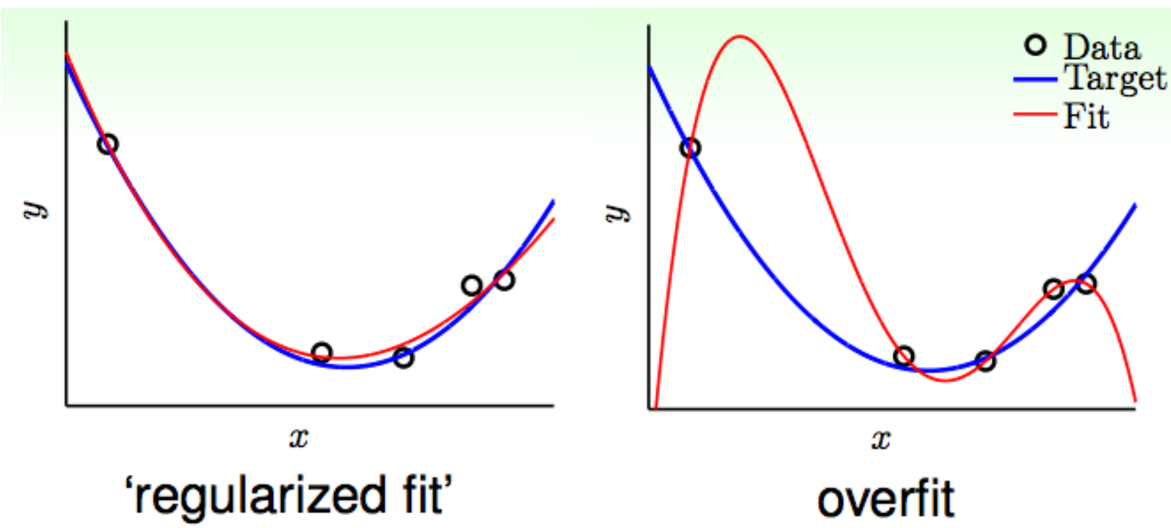
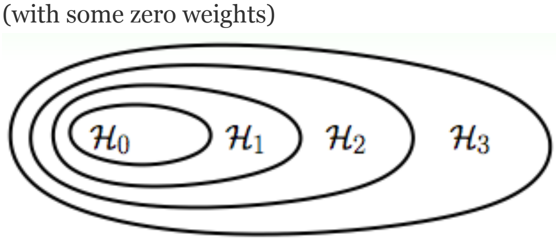
Week 14: Regularization

Table of Contents

Regularized Hypothesis Set



1. Regularization: Function approximation for **ill-posed** problems
    - Force "stepping back" to lower-order hypothesis sets, to alleviate/avoid overfitting when noise is present
    - Recall that lower-order hypothesis sets can be viewed as **subsets** of higher-order hypothesis sets (with some zero weights)



Regression With Constraint

1. Given Q-th order polynomial $transform$ for $x \in \mathbb{R}$:

$$\phi_Q(x) = (1, x, x^2, \ldots, x^Q)\$ + \text{linear regression}$$

   For simplicity, denote the transformed weight $\tilde{w}$ as w
2. If the target function defined above is to be learned by a second-order hypothesis set $\mathcal{H}_2$ and a 10-th order hypothesis set $\mathcal{H}_{10}$, respecttive, there is:

$$\text{hypothesis } w \text{ in} \mathcal{H}_{10} = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \cdots + w_{10} x^{10}$$
$$\text{hypothesis } w \text{ in} \mathcal{H}_{2} = w_0 + w_1 x + w_2 x^2$$

   In other words, hypothesis sets $\mathcal{H}_2$ and $\mathcal{H}_{10}$ are **equivalent under constraint** $w_3 = w_4 = \cdots = w_{10} = 0$
   ⇒ "Stepping back" in hypothesis set is achieved by applying constraints
3. Represent ideas above in terms of optimization objectives:

$$\mathcal{H}_{10} \equiv \left\{ \mathbf{w} \in \mathbb{R}^{10+1} \right\} \qquad \mathcal{H}_2 \equiv \left\{ \mathbf{w} \in \mathbb{R}^{10+1} \text{ while } w_3 = w_4 = \ldots = w_{10} = 0 \right\}$$

   regression with $\mathcal{H}_{10}$:

$$\min_{\mathbf{w} \in \mathbb{R}^{10+1}} E_{\text{in}}(\mathbf{w})$$

   regression with $\mathcal{H}_2$:

$$\min_{\mathbf{w} \in \mathbb{R}^{10+1}} E_{\text{in}}(\mathbf{w}) \quad \text{s.t.} \quad w_3 = w_4 = \ldots = w_{10} = 0$$

   - "Stepping back" in hypothesis set is essentially optimizing E_{in} in the form of a constrained optimization
   - Overkill within the scope of this specific example, but helps illustrating the core idea of regularization
4. Regression with looser constraint

$$\mathcal{H}_2 \equiv \left\{ \mathbf{w} \in \mathbb{R}^{10+1} \text{ while } w_3 = \ldots = w_{10} = 0 \right\} \qquad \mathcal{H}'_2 \equiv \left\{ \mathbf{w} \in \mathbb{R}^{10+1} \text{ while } \geq 8 \text{ of } w_q = 0 \right\}$$

   regression with $\mathcal{H}_2$:

$$\min_{\mathbf{w} \in \mathbb{R}^{10+1}} E_{\text{in}}(\mathbf{w}) \quad \text{s.t.} \quad w_3 = \ldots = w_{10} = 0$$

   regression with $\mathcal{H}'_2$:

$$\min_{\mathbf{w} \in \mathbb{R}^{10+1}} E_{\text{in}}(\mathbf{w}) \quad \text{s.t.} \quad \sum_{q=0}^{10} [\![w_q \neq 0]\!] \leq 3$$

   - Resulting hypothesis $\mathcal{H}'_2$ only requires a **specific number of** w to be zero, not specific ones
   - More flexible than original constrained $\mathcal{H}_1$
   - Less prone to overfitting than $\mathcal{H}_{10}$
   - **Sparse hypothesis set**, NP-hard to solve
5. Regression with softer constraint

$$\mathcal{H}_2' \equiv \left\{ \mathbf{w} \in \mathbb{R}^{10+1} \text{ while } \geq 8 \text{ of } w_q = 0 \right\}$$

regression with $\mathcal{H}_2'$:

$$\min_{\mathbf{w} \in \mathbb{R}^{10+1}} E_{\text{in}}(\mathbf{w}) \text{ s.t. } \sum_{q=0}^{10} [\![w_q \neq 0]\!] \leq 3$$

$$\mathcal{H}(C) \equiv \left\{ \mathbf{w} \in \mathbb{R}^{10+1} \text{ while } \|\mathbf{w}\|^2 \leq C \right\}$$

regression with $\mathcal{H}(C)$:

$$\min_{\mathbf{w} \in \mathbb{R}^{10+1}} E_{\text{in}}(\mathbf{w}) \text{ s.t. } \sum_{q=0}^{10} w_q^2 \leq C$$

- Resulting hypothesis $\mathcal{H}(C)$ has **small weights overall**: $\|w\|^2 \leq C$, where C is a small number
- Does not require a specific number of w to be zero. Instead weights can be any of
  - All w non-zero, but all very small
  - Small number of non-zero w, but *relatively* significant
  - Larger number of non-zero w, but *relatively* small
- $\mathcal{H}(C)$ *overlaps* but not exactly the same as $\mathcal{H}_2'$
- $\mathcal{H}(C)$ provides soft and smooth structure over $C \geq 0$:

$$\mathcal{H}(0) \subset \mathcal{H}(1.126) \subset \cdots \subset \mathcal{H}(1226) \subset \cdots \subset \mathcal{H}(\infty) = \mathcal{H}_{10}$$

  - Contraint essentially non-existent as C approaches infinity

6. Regularized hypothesis set
   - $\mathcal{H}(C)$ is a **regularized hypothesis set**
   - **Regularized hypothesis** w_{REG} is the optiaml solution from $\mathcal{H}(C)$

Weight Decay Regularization

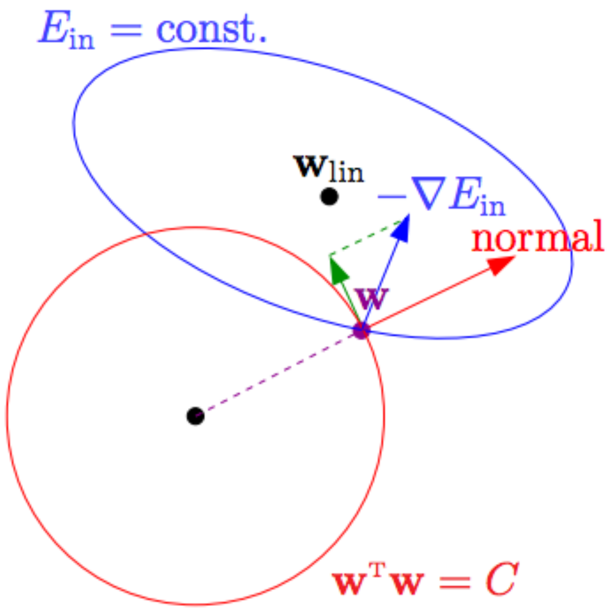
1. Matrix form of regularized regression problem

$$\min_{w \in \mathbb{R}^{Q+1}} E_{in}(w) = \frac{1}{N} \underbrace{\sum_{n=1}^{N} (w^T z_n - y_n)^2}_{(Zw - y)^T (Zw - y)}$$

$$\text{s.t.} \quad \underbrace{\sum_{q=0}^{Q} w_q^2}_{w^T w} \leq C$$

   - Constraint $w^T w \leq C$ requires that feasible w resides within a **radius-$\sqrt{C}$ hypersphere**

Lagrange Multiplier



1. Target function

$$\min_{w \in \mathbb{R}^{Q+1}} E_{in}(w) = \frac{1}{N}(Zw - y)^T(Zw - y) \text{ s.t. } w^T w \leq C$$

2. Guiding principle for minimizing E_{in}: Gradient descent in the direction of $-\nabla E_{in}(w)$
3. In order to satisfy the constraint while seeking optimal solution, need to make sure that w **does not take value outside of the bound indicated by red circle above**
   - Candidate w are most likely located along the **boundary** defined by $w^T w = C$, since it allows w to be as close to **unconstrained optimal** w_{LIN} as possible
   - Denote *normal* vector at the boundary of the constraint circle as $\vec{w}$
   - **Cannot** iterate along direction of negative gradient if $-\nabla E_{in} \parallel \vec{w}$
     - Otherwise w violates the constraint upon next iteration
   - If $-\nabla E_{in}$ and w are **not** parallel with each other, **it is possible to decrease $E_{in}(w)$ without violating the constraint**
     - There exists a component vector (denoted in green in picture above) **along** the constraint boundary, which allows w to move closer (with infinitely small step size) to w_{LIN} **without** moving out of the boundary
4. Optimal \color{purple}(w_{REG}) must satisfy:
   - Gradient $-\nabla E_{in}$ at w_{REG} is **parallel** with the normal vector of $w^T w = C$ at w_{REG}
     - Otherwise further optimization is possible
     - In this case, the normal vector is w_{REG} itself (denoted as $\boxed{w_{REG}}$), hence:

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

5. To solve for optimum *regularized* weight $\boxed{w_{REG}}$, need to find Lagrange multiplier $\lambda > 0$ such that

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

   - The extra constant $\frac{2}{N}$ simplifies the solution, without impacting the optimal w_{REG}, because w_{REG} can always be treated as a unit vector

Augmented Error

1. Given $\lambda > 0$, there is

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

   Recall the unconstrained optimal w from linear regression

   $$\frac{2}{N}(Z^T Z w_{REG} - Z^T y) + \frac{2\lambda}{N}\boxed{w_{REG}} = 0$$
   $$(Z^T Z w_{REG} - Z^T y) + \lambda\boxed{w_{REG}} = 0$$
   $$(Z^T Z + \lambda) w_{REG} = Z^T y$$

   Recall that $Z^T Z$ is semi-positive definite. LHS is therefore positive definite (invertible) since λ is assumed to be p

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

   - The process for solving regularized optimal weight w_{REG}, as shown above, is known as **ridge regression** in statistics, a.k.a **weight decay** in machine learning

2. Augmented error

   - Generalize the solution above beyond linear regression cases
   - Solving the gradient equation

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

     Is equivalent to minimizing the integral form

     $$E_i n(w) + \underbrace{\frac{\lambda}{N}\overbrace{w^T w}^{regularizer}}_{\text{augmented error } E_{aug}(w)}$$

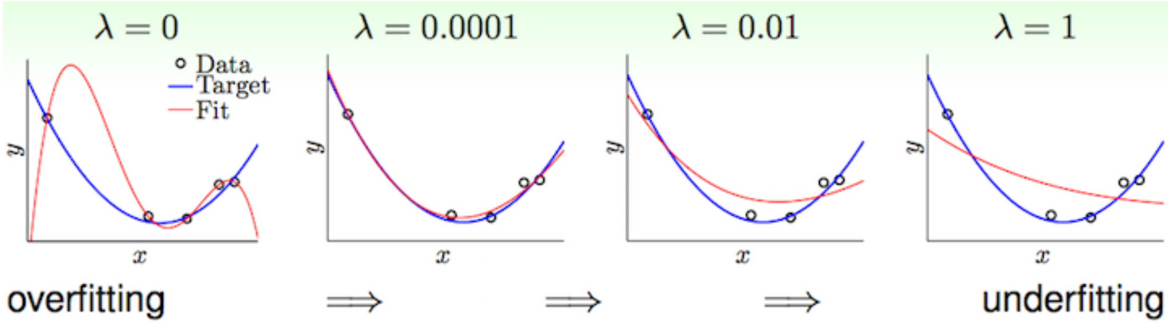
     Recall that $w^T w$ is the matrix form of w_{REG}^2

3. Given $\lambda > 0$, the constrained regression problem essentially becomes regularization with augmented error instead of *constrained* E_{in}

   - Minimizing *unconstrained* E_{aug} effectively minimizes some $C - constrained$ E_{in}
   - In the special case of $\lambda = 0$, it becomes normal unconstrained regression problem, and $E_{aug} = E_{in}$

   $$w_{REG} \leftarrow \arg\min_{w} E_{aug}(w) \text{ for given } \lambda > 0 \text{ or } \lambda = 0$$

4. Effect of regularization



- A little **regularization** goes a long way!

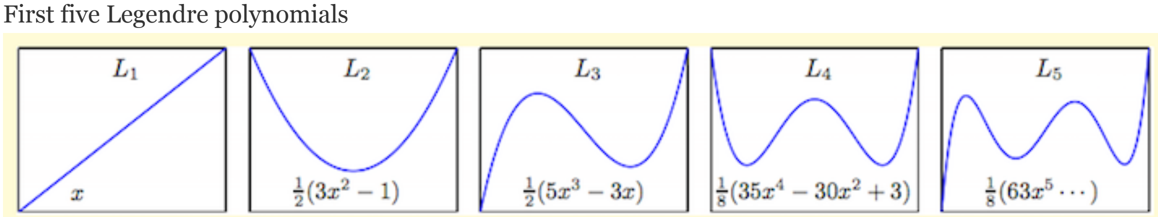
5. The term $+ \frac{\lambda}{N} w^T w$ is known as **weight-decay** regularization
   - Larger $\lambda \iff$ prefer shorter $w \iff$ effectively smaller C (tighter constraint)
   - Works with any transform + linear model

Legendre Polynomials

1. General optimization problem for regularized regression with non-linear transformation

$$\min_{w \in \mathbb{R}^{Q+1}} \frac{1}{N} \sum_{n=0}^{N} (w^T \phi x_n - y_n)^2 + \frac{\lambda}{N} \sum_{q=0}^{Q} w_q^2$$

2. Problem with naive polynomial transform: $\phi(x) = (1, x, x^2, \ldots, x^Q)$
   - When x_n is small ($x_n \in [-1, +1]$), higher-order x_n^q is very small to begin with and requires very large w_q to actually cause overfitting
   - Regularization might **over-punish** higher-order terms in this case.
3. Use *normalized* polynomial transform: $(1, L_1(x), L_2(x), \ldots, L_Q(x))$ to avoid over-regularization
   - Treating polynomial terms as vectors
   - Require that the inner products of these vector terms to be small (or zero)
   - **Orthonormal basis functions**: [Wikipedia]
   - Known as **Legendre polynomials**
4. Using Legendre polynomials in place of naive polynomial transformations can help produce better results when using regularizations in polynomial regressions
5. First five Legendre polynomials



Regularization and VC Theory

1. VC guarantee of regularized regressions

Regularization by Constrained-Minimizing E_{in}	Regularization by Minimizing E_{aug}	VC Guarantee of Constrained-Minimizing E_{in}
$\min_{\mathbf{w}} E_{\text{in}}(\mathbf{w})$ s.t. $\mathbf{w}^T\mathbf{w} \le C$	$\min_{\mathbf{w}} E_{\text{aug}}(\mathbf{w}) = E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w}^T\mathbf{w}$	$E_{\text{out}}(\mathbf{w}) \le E_{\text{in}}(\mathbf{w}) + \Omega(\mathcal{H}(C))$

- Constrained-minimizing E_{in} $\overset{C \text{ equivalent to some } \lambda}{\Longleftrightarrow}$ Minimizing (unconstrained) E_{aug}
- Constrained-minimizing E_{in} $\overset{\text{provides}}{\Longleftrightarrow}$ VC guarantee (under constrained hypothesis set $\mathcal{H}(C)$)
- Given the equivalence relationship between E_{in} and E_{aug}, solution to augmented error problem provides the **same VC guarantee without the hypothesis set constraint**

2. Another view of augmented error

Augmented Error	VC Bound
$E_{\text{aug}}(\mathbf{w}) = E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w}^T\mathbf{w}$	$E_{\text{out}}(\mathbf{w}) \leq E_{\text{in}}(\mathbf{w}) + \Omega(\mathcal{H})$

- Regularizer $w^T w$: Penalizing complexity of **a single hypothesis** (specified by value of w)
- Generalization price $\Omega(\mathcal{H})$: Penalizing complexity of **a hypothesis set**
- Given the similarity between regularizer and generalization price, if $\frac{\lambda}{N}$ $\Omega(w)$ is a good representation of $\Omega(\mathcal{H})$, E_{aug} is a **better proxy** of E_{out} than E_{in}

3. Minimizing E_{aug}

- (Heuristically) allows learning algorithm to operate with better proxy of E_{out}
- (Technically) allows learning algorithm to enjoy flexibility of the whole hypothesis set $\mathcal{H}$

4. Effective VC dimension

- When minimizing augmented error

$$\min_{w \in \mathbb{R}^{\tilde{d}+1}} E_{aug}(w) = E_{in}(w) + \frac{\lambda}{N}\Omega(w)$$

- Model complexity $d_{VC}(\mathcal{H}) = \tilde{d} + 1$, because all possible w are considered during minimization
- However, only $\mathcal{H}(C)$ choices of w are **actually considered**, with some C equivalent to λ
  - Unconstrained minimization, but accounting for constraints in target function
- The effective VC dimension is therefore **smaller** than that obtained from solving unconstrained E_{in}

$$d_{VC}(\mathcal{H}(C)) = d_{EFF}(\mathcal{H}, \underbrace{\mathcal{A}}_{\min E_{aug}})$$

- Depending on the complexity of original hypothesis set,original VC dimension $d_{VC}(\mathcal{H})$ could be large. However, the **effective** VC dimension $d_{EFF}(\mathcal{H}, \mathcal{A})$ can remain small if $\mathcal{A}$ is regularized.

General Regularizers <a name="general-regularizers"/ >

1. General guideline for choosing general regularizers Ωw for target function:

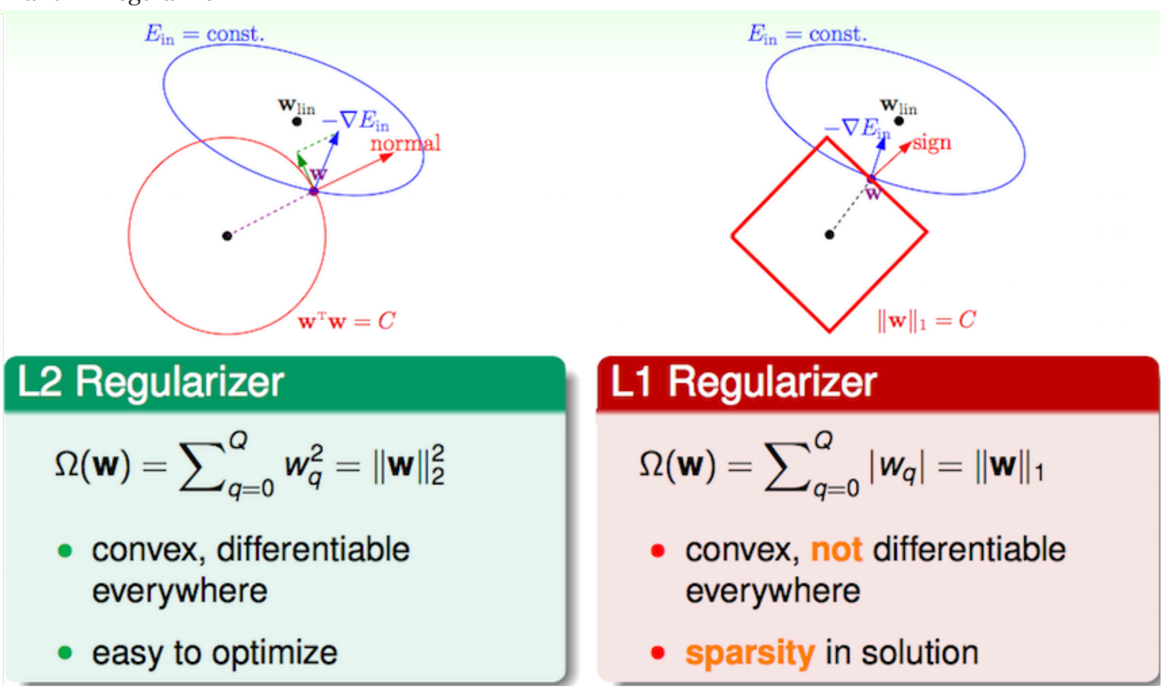
- **Target-dependent**: Make use of *properties* of target function, if known
  - The main purpose of regularizers is to get learning results closer to target function
  - e.g. If target function is known two be an even function, use symmetric regularizer: $\sum \|q \text{ is odd}\| w_q^2$
- **Plausible**: Direction towards smoother or simpler hypothesis
  - Stochastic/deterministic noise are both **non-smooth**
  - Sparsity(L1) regularizer: $\sum |w_q|$
- **Friendly**: Easy to optimize

- Weight-decay(L1) regularizer: $\sum w^2$
- **No regularizer**: $\lambda = 0$, always an option

2. Connection between regularizer and error measure

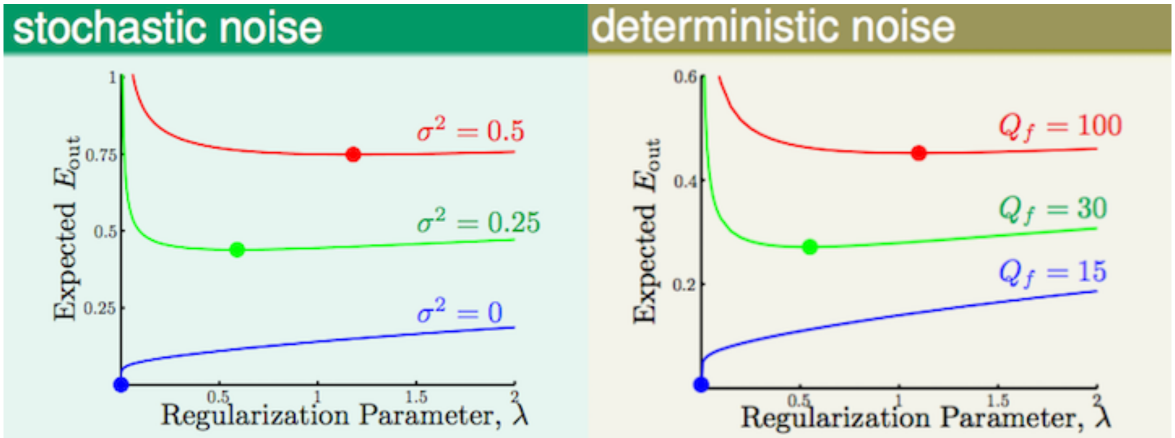
   - augmented error = error $\hat{err}$ + regularizer Ω
   - Given the relationship above, the guidelines for choosing regularizers and error measures share some commonalities
     - Regularizer: target-dependent, plausible, *or* friendly
     - Error measure: user-dependent, plausible, *or* friendly

3. L1 and L2 regularizer



   - L1 regularizer tends to produce sparse solutions (only a few non-zero weights) because given the "square" nature of its constraint boundary, there is a tendency for the optimal solutions to be at one of the corners.
   - L1 regularizer is therefore suitable for feature selection (zero-out less important feature) or use cases that require sparse solution for computational simplicity
   - Regression that uses L1 regularization is called **Lasso Regression** (Least Absolute Shrinkage and Selection Operator)
   - Regression that uses L2 regularization is called **Ridge Regression**

4. Choosing the optimal λ



- Annotations
  - σ^2 : Amount of stochastic noise
  - Q_f : Order of target function (deterministic noise is the delta between this value and that of the hypothesis set, assume to be at 15-th order)
- More noise, more regularization required to achieve smooth solution with small E_{out}