شبکه‌های عصبی مصنوعی

درس ۱۳

تعمیم

Generalization

کاظم فولادی قلعه

دانشکده مهندسی، پردیس فارابی

دانشگاه تهران

http://courses.fouladi.ir/nn

Generalization

تعمیم

GENERALIZATION

یکی از موارد کلیدی در طراحی MLP تعیین تعداد نرون‌های مورد استفاده است:

اگر تعداد نرون‌ها خیلی زیاد باشد ⇐ شبکه بر روی داده‌های آموزشی بیش‌برازش (overfit) می‌کند.
(**بیش‌برازش:** خطا بر روی داده‌های آموزشی بسیار کوچک ـــ اما ـــ خطا در برابر داده‌های جدید بسیار بزرگ)

شبکه‌ای که **تعمیم** آن خوب باشد، بر روی داده‌های جدید به‌خوبی داده‌های آموزشی عمل می‌کند.

احتمال بیش‌برازش و تعمیم پایین در یک شبکه‌ی پیچیده، بالاتر است.

پیچیدگی یک شبکه‌ی عصبی بر اساس تعداد پارامترهای آزاد آن (وزن‌ها و بایاس‌ها) مشخص می‌شود
(تعداد پارامترهای آزاد تابعی است از تعداد نرون‌ها).

هدف این فصل:
«تنظیم پیچیدگی شبکه به‌منظور متناسب شدن با پیچیدگی داده‌ها»
این کار می‌تواند بدون تغییر تعداد نرون‌ها انجام شود:
می‌توانیم تعداد پارامترهای آزاد مؤثر را بدون تغییر تعداد پارامترهای آزاد واقعی تنظیم کنیم.

Generalization

A cat that once sat on a hot stove
will never again sit on a hot stove
or on a cold one either.

Mark Twain

تعمیم

Generalization

“A cat that once sat on a hot stove will never again sit on a hot stove or on a cold one either.”

Mark Twain

«گربه‌ای که یک بار روی بخاری داغ نشسته باشد، هرگز دوباره روی یک بخاری داغ و یا حتی سرد نخواهد نشست.»
مارک تواین

«مارگزیده از ریسمان سیاه و سفید می‌ترسد!»

تعمیم

۱

صورت مسئله

Generalization

- The network input-output mapping is accurate for the training data and for test data never seen before.

- The network interpolates well.

تعمیم

GENERALIZATION

منظور از تعمیم:

- The network input-output mapping is accurate for the training data and for test data never seen before.
- The network interpolates well.

- نگاشت ورودی ـ خروجی برای داده‌های آموزشی و برای داده‌های آزمایشی که پیش از این هرگز دیده نشده‌اند، دقیق باشد.
- شبکه به‌خوبی درون‌یابی می‌کند.

Cause of Overfitting

Poor generalization is caused by using a network that is too complex (too many neurons/parameters). To have the best performance we need to find the least complex network that can represent the data (Ockham's Razor).

علت بیش‌برازش

CAUSE OF OVERFITTING

تعمیم ضعیف در اثر استفاده از یک شبکه‌ی بسیار پیچیده (تعداد زیادی نرون/ پارامتر) ناشی می‌شود.

برای داشتن بهترین کارآیی، لازم است شبکه‌ای با حداقل پیچیدگی را بیابیم که بتواند داده‌ها را بازنمایی کند (**تیغه‌ی اوخامی**).

Poor generalization is caused by using a network that is too complex (too many neurons/parameters).
To have the best performance we need to find the least complex network that can represent the data (Ockham’s Razor).

Ockham's Razor

Find the simplest model that explains the data.

تیغه‌ی اوخامی

OCKHAM'S RAZOR

اصل تیغه‌ی اوخامی
Ockham's Razor

ساده‌ترین مدلی که داده‌ها را توضیح می‌دهد، بیابید.

Find the simplest model that explains the data.

هرچه مدل پیچیده‌تر باشد، امکان خطا بیشتر می‌شود.

Problem Statement

Training Set

$$\{\mathbf{p}_1, \mathbf{t}_1\}\ ,\ \{\mathbf{p}_2, \mathbf{t}_2\}\ ,\ \dots\ ,\ \{\mathbf{p}_Q, \mathbf{t}_Q\}$$

Underlying Function

$$\mathbf{t}_q = \mathbf{g}(\mathbf{p}_q) + \varepsilon_q$$

Performance Function

$$F(\mathbf{x}) = E_D = \sum_{q=1}^{Q} (\mathbf{t}_q - \mathbf{a}_q)^T (\mathbf{t}_q - \mathbf{a}_q)$$

تعمیم

صورت مسئله

PROBLEM STATEMENT

Training Set

$$\{\mathbf{p}_1, \mathbf{t}_1\}\ ,\ \{\mathbf{p}_2, \mathbf{t}_2\}\ ,\ \dots\ ,\ \{\mathbf{p}_Q, \mathbf{t}_Q\}$$

Underlying Function

$$\mathbf{t}_q = \mathbf{g}(\mathbf{p}_q) + \varepsilon_q$$

فرض می‌کنیم تارگت‌ها با این رابطه ساخته شده‌اند:
$\mathbf{g}(.)$: یک تابع مجهول
ε_q : یک نویز تصادفی مستقل با میانگین صفر

هدف آموزش: یافتن یک شبکه‌ی عصبی برای تقریب $\mathbf{g}(.)$ با نادیده گرفتن نویز

Performance Function

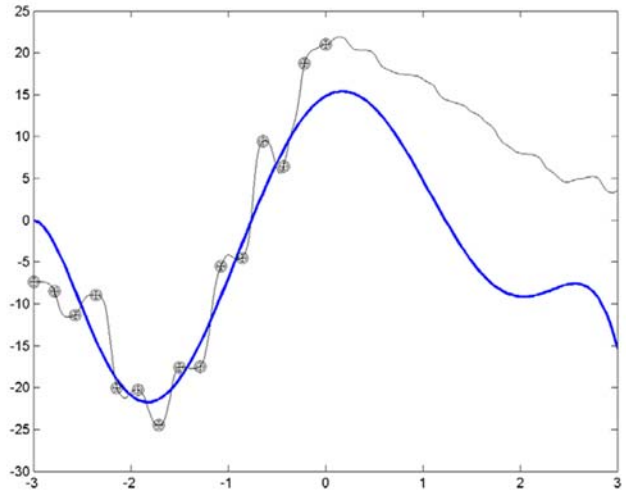
$$F(\mathbf{x}) = E_D = \sum_{q=1}^{Q} (\mathbf{t}_q - \mathbf{a}_q)^T (\mathbf{t}_q - \mathbf{a}_q)$$

SSE روی داده‌های آموزشی D

شاخص کارآیی استاندارد
برای آموزش شبکه‌ی عصبی:
مجموع مربعات خطا (SSE)

Poor Generalization

Overfitting Extrapolation



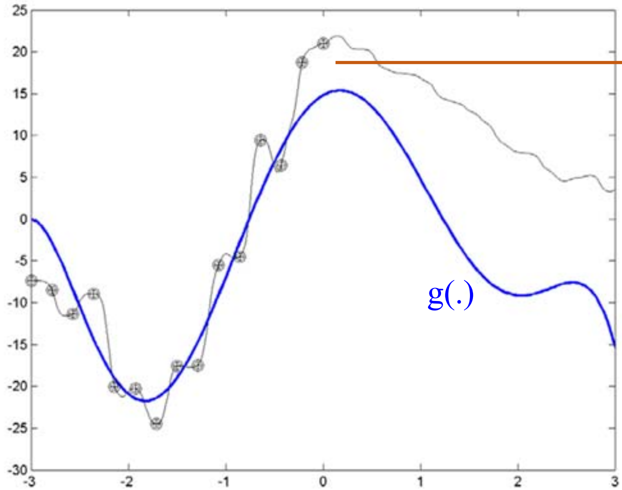
Interpolation

تعمیم

تعمیم ضعیف

POOR GENERALIZATION

بیش‌برازش
Overfitting
برون‌یابی
Extrapolation



در این ناحیه دادهی آموزشی وجود ندارد.

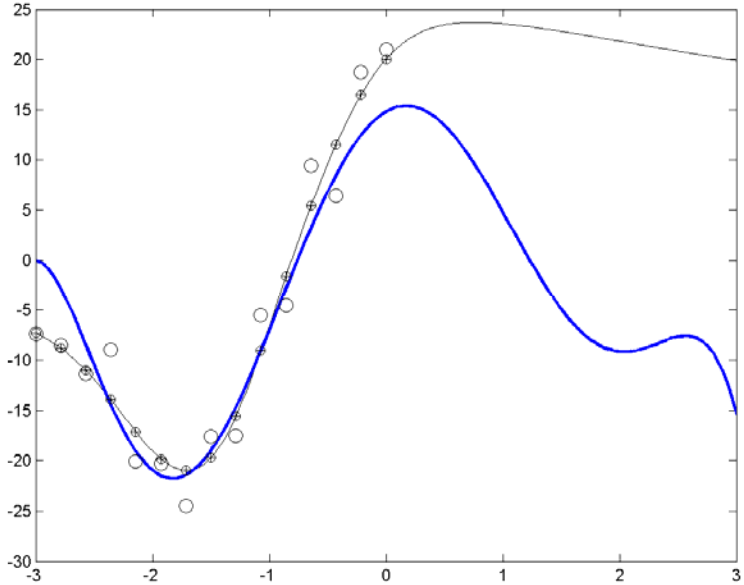
Interpolation
درون‌یابی

دو نوع خطا
- در درون‌یابی
- در برون‌یابی

هدف اصلی: جلوگیری از خطاهای **درون‌یابی**
راهی برای جلوگیری از خطاهای **برون‌یابی** وجود ندارد،
مگر اینکه داده‌های آموزشی کل ناحیه‌ی مورد استفاده در شبکه را پوشش دهد.

Good Generalization

Interpolation Extrapolation



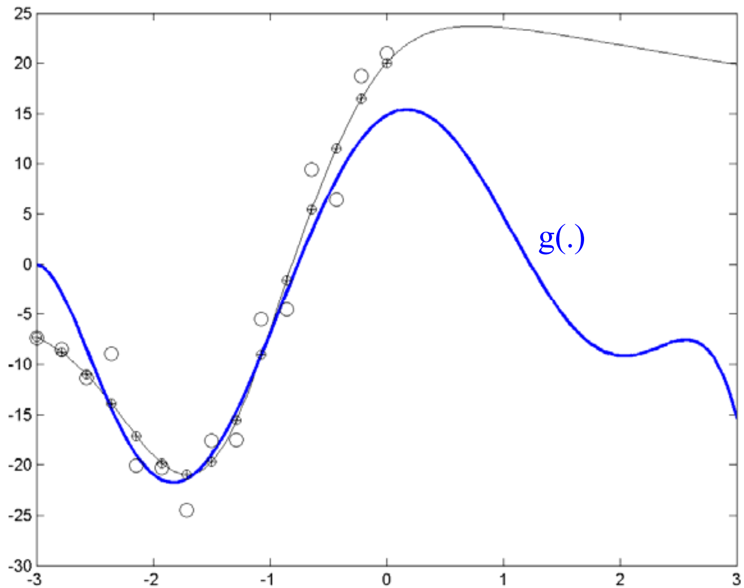
تعمیم

تعمیم ضعیف

Poor Generalization

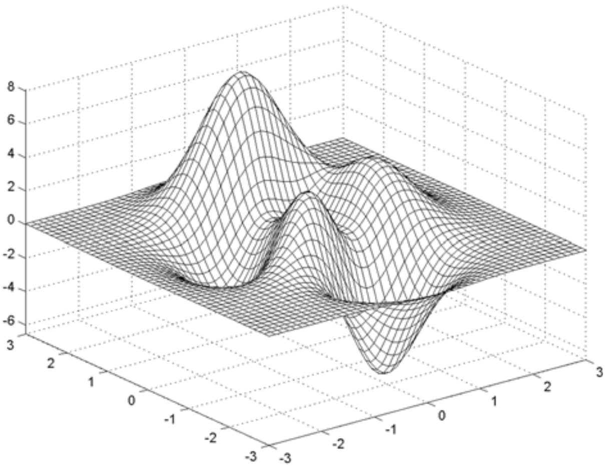
درون‌یابی
Interpolation

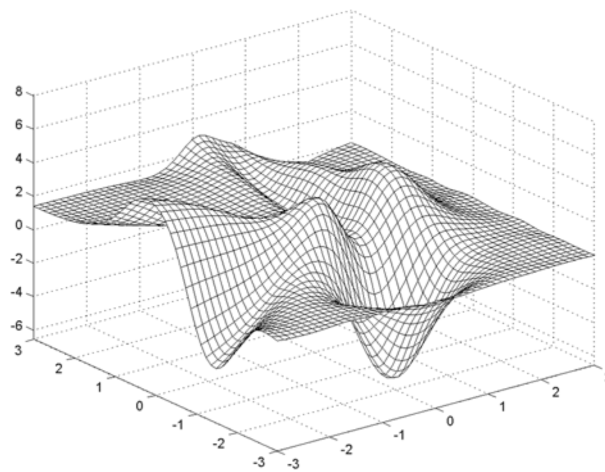
برون‌یابی
Extrapolation



شبکه‌ای مشابه مثال قبل با همان تعداد وزن‌ها و همان داده‌ها (اما بدون استفاده از همه‌ی وزن‌های موجود).
پاسخ شبکه کاملاً بر روی تابع تطابق ندارد، اما کاری که می‌کند بهترین کاری است که بر اساس داده‌های محدود نویزی می‌تواند.

Extrapolation in 2-D

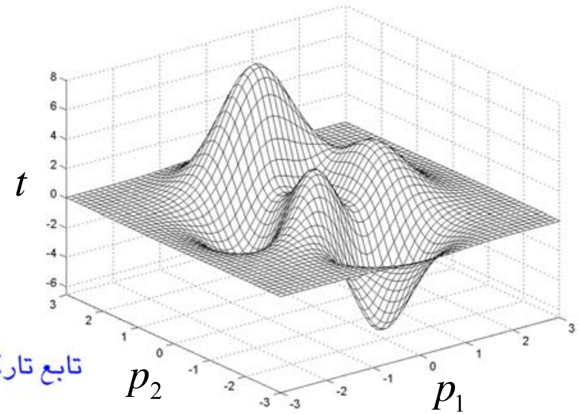




تعمیم

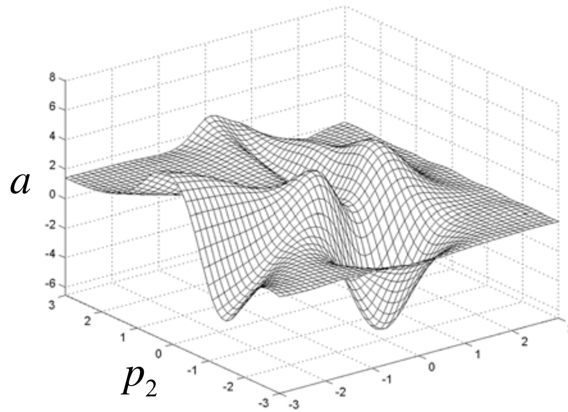
برون‌یابی در دو بعد

EXTRAPOLATION IN 2-D



تابع تارگت

تقریب توسط شبکه‌ی عصبی



وقتی شبکه تعداد زیادی ورودی دارد،
تعیین اینکه چه زمانی درون‌یابی و چه زمانی برون‌یابی می‌کند، دشوارتر است.

تعمیم‌پذیر کردن شبکه‌های عصبی

برای ایجاد تعمیم‌پذیری، روی‌کرد کلی سعی در یافتن ساده‌ترین شبکه است که بر داده‌ها fit شود.

دو دسته‌ی کلی از روش‌های تعمیم‌پذیرسازی

۱) محدود کردن تعداد وزن‌ها (تعداد نرون‌ها)

۲) محدود کردن اندازه‌ی وزن‌ها

توقف زودهنگام
Early Stopping

رگولاریزاسیون
Regularization

شبکه های عصبی مصنوعی

تعمیم

۲

تخمین خطای تعمیم

Measuring Generalization

Test Set

- Part of the available data is set aside during the training process.
- After training, the network error on the test set is used as a measure of generalization ability.
- The test set must never be used in any way to train the network, or even to select one network from a group of candidate networks.
- The test set must be representative of all situations for which the network will be used.

تخمین خطای تعمیم برای یک شبکه‌ی عصبی خاص

اندازه‌گیری تعمیم

MEASURING GENERALIZATION

مجموعه‌ی آزمایشی
Test Set

- بخشی از داده‌های موجود در طول فرآیند آموزش کنار گذاشته می‌شوند.
- پس از آموزش، خطای شبکه بر روی مجموعه‌ی آزمایشی به‌عنوان معیار تعمیم‌پذیری استفاده می‌شود.

دو ویژگی مهم برای مجموعه‌ی آزمایشی

(برای اینکه شاخص معتبری برای تعمیم‌پذیری باشد):

- مجموعه‌ی آزمایشی نباید به‌هیچ وجه برای آموزش شبکه استفاده شود،
یا حتی برای انتخاب یک شبکه از یک گروه از شبکه‌های کاندیدا به‌کار برده شود.
- مجموعه‌ی آزمایشی باید نماینده‌ی همه‌ی موقعیت‌ها برای مواردی باشد که شبکه در آنها به‌کار خواهد رفت.
(تضمین این شرط بسیار دشوار است، به‌خصوص وقتی فضای ورودی دارای بعد بالا باشد یا شکل پیچیده داشته باشد)

تذکر:

فرض بر این است که مقدار داده‌ها (داده‌های آموزشی شبکه) **محدود** است.
اگر مقدار داده‌ها نامحدود باشد، در این صورت مشکل **بیش‌برازش** وجود نخواهد داشت.
(نامحدود به‌لحاظ عملی: تعداد نقاط داده به طور قابل توجهی بزرگ‌تر از تعداد پارامترهای شبکه باشد)

تعمیم

۳

روش‌هایی برای بهبود تعمیم

Methods for Improving Generalization

- Pruning (removing neurons) until the performance is degraded.
- Growing (adding neurons) until the performance is adequate.
- Validation Methods
- Regularization

روش‌هایی برای بهبود تعمیم

METHODS FOR IMPROVING GENERALIZATION

رشد شبکه *Network Growing*	شروع از یک شبکه بدون نرون و افزودن نرون تا رسیدن به کارآیی کافی
هرس شبکه *Network Pruning*	شروع از یک شبکه‌ی بزرگ (احتمالاً با بیش‌برازش) و حذف یک به یک نرون‌ها تا کاهش چشمگیر کارآیی
جستجوهای سراسری *Global Searches*	استفاده از روش‌هایی مثل الگوریتم ژنتیک برای جستجو در همه‌ی معماری‌های ممکن و انتخاب بهترین آنها
روش‌های اعتبارسنجی *Validation Methods*	استفاده از یک مجموعه داده علاوه بر داده‌های آموزش و آزمایش برای اعتبارسنجی و مثلاً توقف زودهنگام (early stopping)
رگولاریزاسیون *Regularization*	کوچک نگاه داشتن شبکه با مقید کردن بزرگی وزن‌های آن (به‌جای مقید کردن تعداد وزن‌ها)

تعمیم

۴

توقف زودهنگام

13

Early Stopping

- Break up data into training, *validation*, and test sets.
- Use only the training set to compute gradients and determine weight updates.
- Compute the performance on the validation set at each iteration of training.
- Stop training when the performance on the validation set goes up for a specified number of iterations.
- Use the weights which achieved the lowest error on the validation set.

توقف زودهنگام

منطق

EARLY STOPPING

توقف زودهنگام، ساده‌ترین روش تعمیم‌پذیر کردن است:

ایده: با پیشرفت فرآیند آموزش، شبکه بیشتر و بیشتراز وزن‌هایش استفاده می‌کند تا همه‌ی وزن‌ها استفاده شوند.

- وقتی آموزش به می‌نیمم رویه‌ی خطا رسید، با افزایش تعداد تکرارهای آموزش، پیچیدگی شبکه‌ی حاصل افزایش می‌یابد.
- اگر آموزش پیش از رسیدن به می‌نیمم متوقف شود، شبکه به‌طور مؤثر از تعداد پارامتر کمتری استفاده خواهد کرد و احتمال بیش‌برازش در آن کمتر است.

توقف زودهنگام

روش

EARLY STOPPING

مجموعه‌ی داده‌ها را به سه زیرمجموعه تقسیم می‌کنیم:

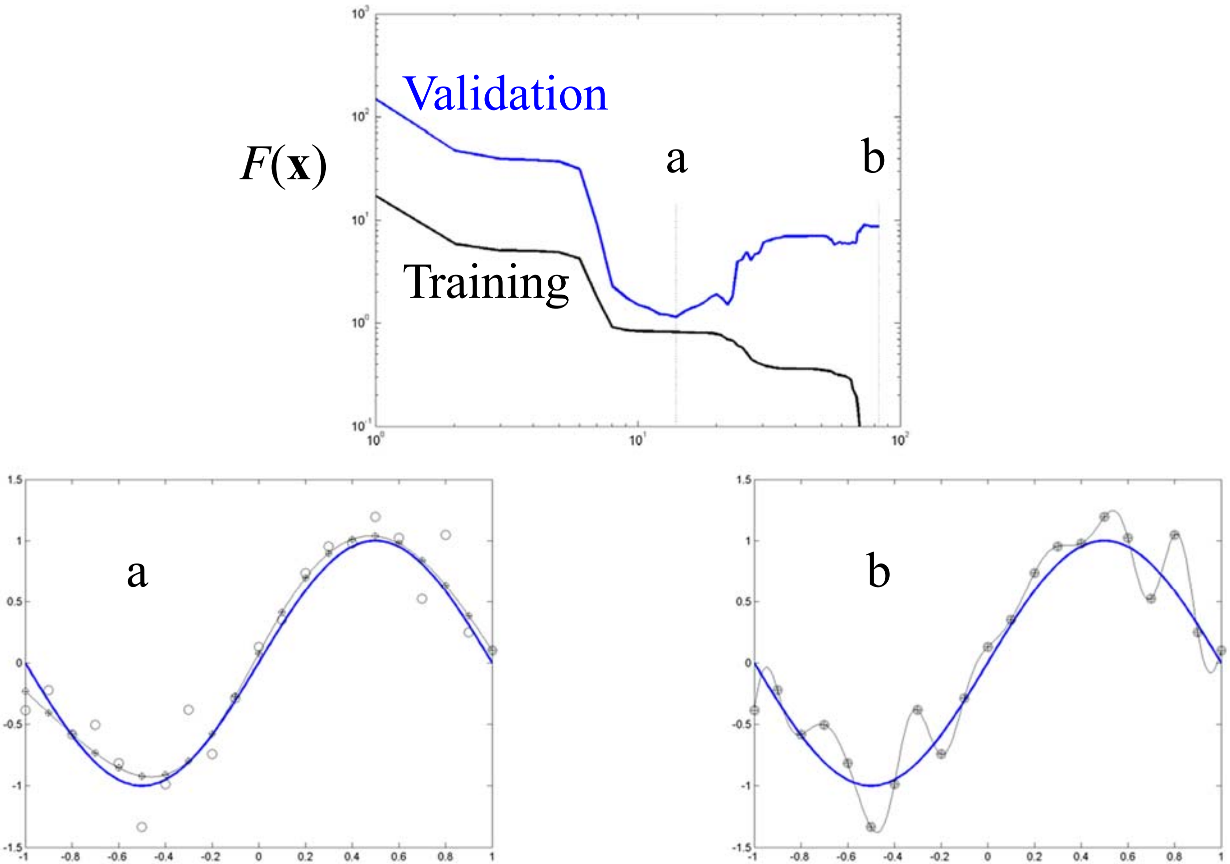
مجموعه‌ی داده‌ها		
مجموعه‌ی آموزشی *Training Set*	مجموعه‌ی اعتبارسنجی *Validation Set*	مجموعه‌ی آزمایشی *Test Set*

- Use only the training set to compute gradients and determine weight updates.
- Compute the performance on the validation set at each iteration of training.
- Stop training when the performance on the validation set goes up for a specified number of iterations.
- Use the weights which achieved the lowest error on the validation set.

- برای محاسبه‌ی وزن‌ها (به‌هنگام‌سازی) و محاسبه‌ی گرادیان‌ها فقط از مجموعه‌ی آموزشی استفاده می‌کنیم.
- در هر تکرار آموزش، کارآیی را بر روی مجموعه‌ی اعتبارسنجی محاسبه می‌کنیم.
- وقتی کارآیی بر روی مجموعه‌ی اعتبارسنجی طی تعدادی تکرار بالا رفت (بدتر شد)، آموزش را متوقف می‌کنیم.
- از وزن‌هایی که پایین‌ترین میزان خطا روی مجموعه‌ی اعتبارسنجی را به‌دست می‌دهند، استفاده می‌کنیم.

چه زمانی باید آموزش را متوقف کنیم؟ استفاده از cross-validation

Early Stopping Example



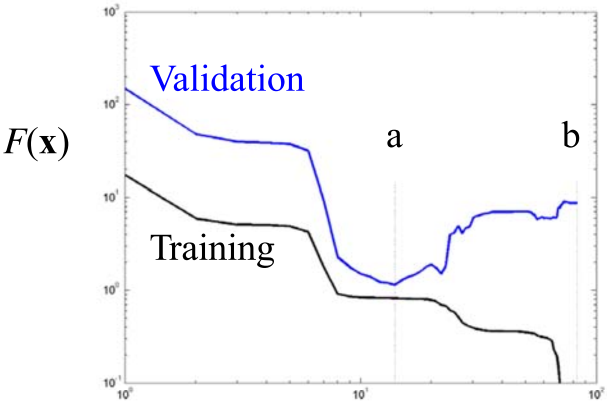
توقف زودهنگام

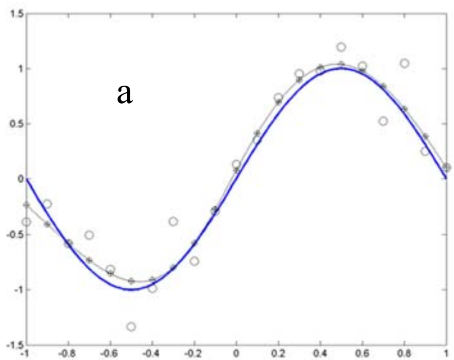
مثال

EARLY STOPPING EXAMPLE

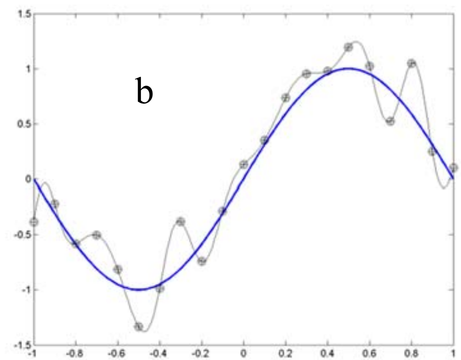
مجموعه‌ی داده‌ها		
مجموعه‌ی آموزشی *Training Set*	مجموعه‌ی اعتبارسنجی *Validation Set*	مجموعه‌ی آزمایشی *Test Set*
70%	**15%**	**15%**

- هر سه مجموعه باید نماینده‌ی خوبی از کل فضای پارامترها باشند (با اندازه‌های مختلف).
- در این روش باید از یک الگوریتم آموزش نسبتاً کند استفاده کرد (اگر روش خیلی سریع باشد احتمالاً به‌سرعت به نقطه‌ی می‌نیمم خطای اعتبارسنجی میل می‌کند).

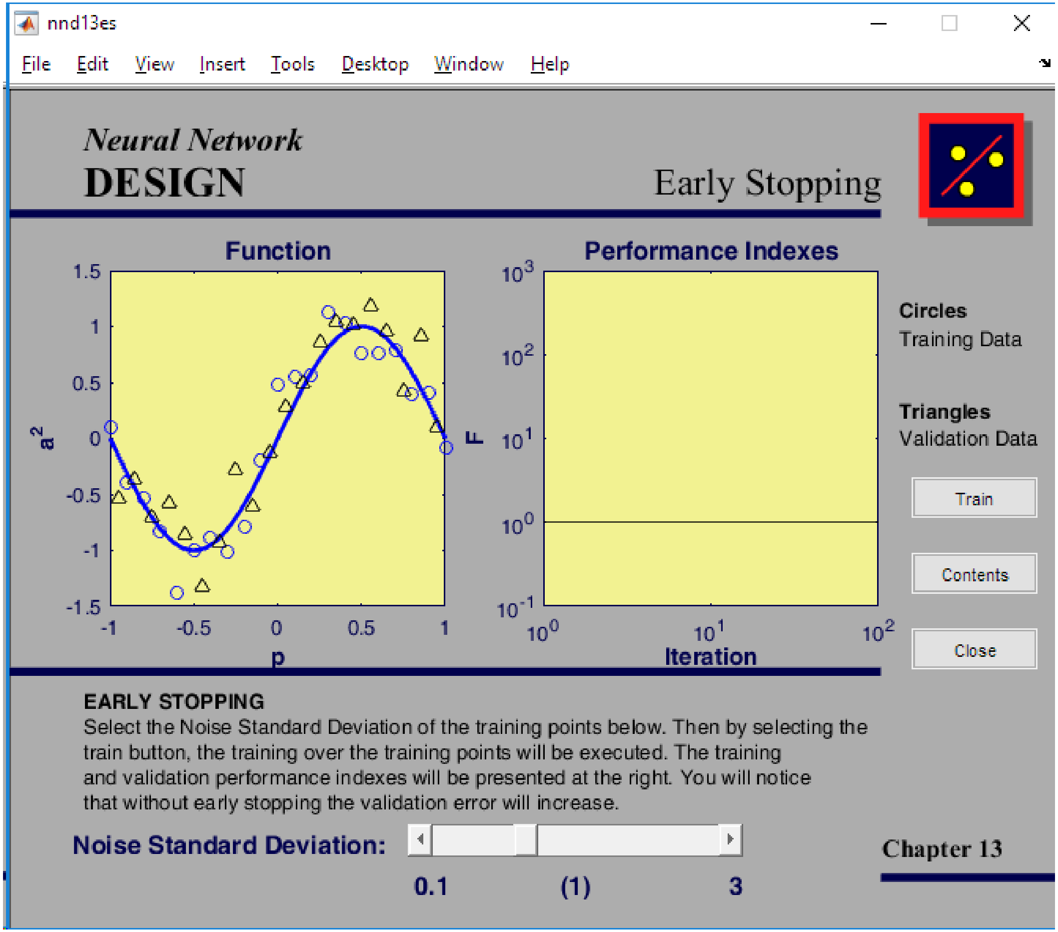




تعمیم خوب



تعمیم بد


nnd13es
File Edit View Insert Tools Desktop Window Help
Neural Network
DESIGN
Early Stopping
Function
Performance Indexes
Circles
Training Data
Triangles
Validation Data
Train
Contents
Close
p
Iteration
EARLY STOPPING
Select the Noise Standard Deviation of the training points below. Then by selecting the train button, the training over the training points will be executed. The training and validation performance indexes will be presented at the right. You will notice that without early stopping the validation error will increase.
Noise Standard Deviation:
0.1
(1)
3
Chapter 13

```
>> nnd13es
```

تعمیم

۵

رگولاریزاسیون
(تنظیم)

Regularization

Standard Performance Measure

$$F = E_D$$

Performance Measure with Regularization

$$F = \beta E_D + \alpha E_W = \beta \sum_{q=1}^{Q} (\mathbf{t}_q - \mathbf{a}_q)^T (\mathbf{t}_q - \mathbf{a}_q) + \alpha \sum_{i=1}^{n} x_i^2$$

Complexity Penalty

)Smaller weights means a smoother function(.

رگولاریزاسیون

REGULARIZATION

در رگولاریزاسیون، شاخص کارآیی SSE را تغییر می‌دهیم تا شامل جمله‌ای برای جریمه کردن پیچیدگی شبکه شود.

Standard Performance Measure

$$F = E_D$$

Performance Measure with Regularization

جمله‌ای شامل مشتقات یک تابع تقریبی وارد می‌شود که تابع حاصل را مجبور می‌کند «هموار: smooth» شود.

$$F = \beta E_D + \alpha E_W = \beta \sum_{q=1}^{Q} (\mathbf{t}_q - \mathbf{a}_q)^T (\mathbf{t}_q - \mathbf{a}_q) + \alpha \sum_{i=1}^{n} x_i^2$$

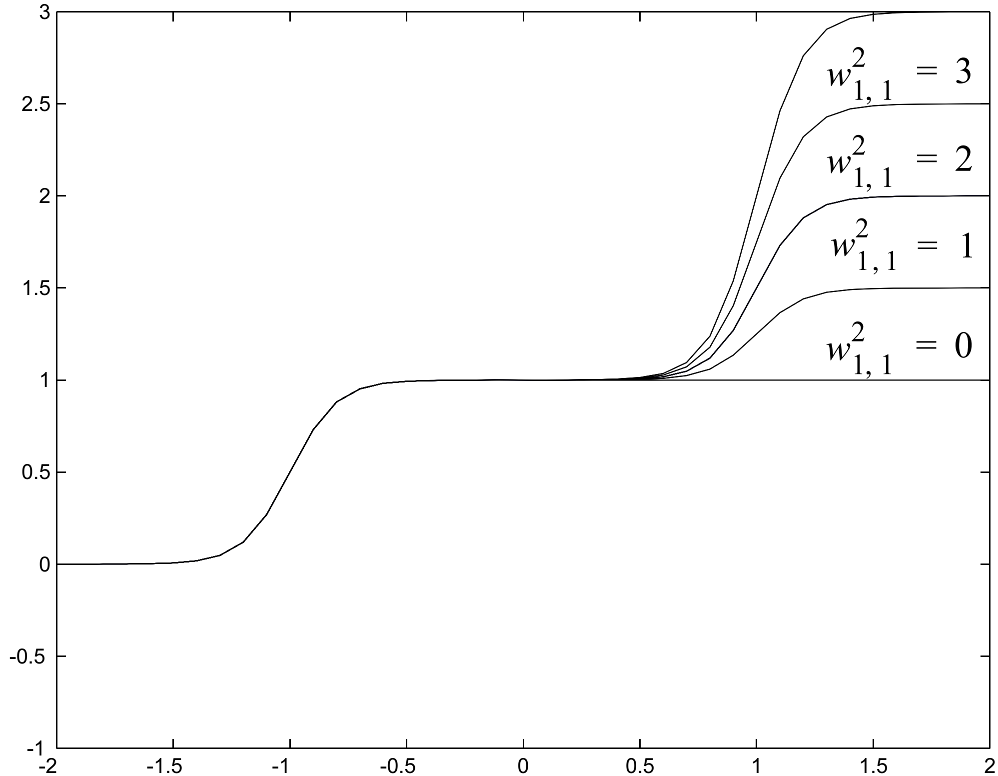
Complexity Penalty

مجموع مربعات وزن‌های شبکه

(Smaller weights means a smoother function.)

هر چه نسبت α/β بزرگتر باشد، پاسخ شبکه هموارتر خواهد بود.

Effect of Weight Changes

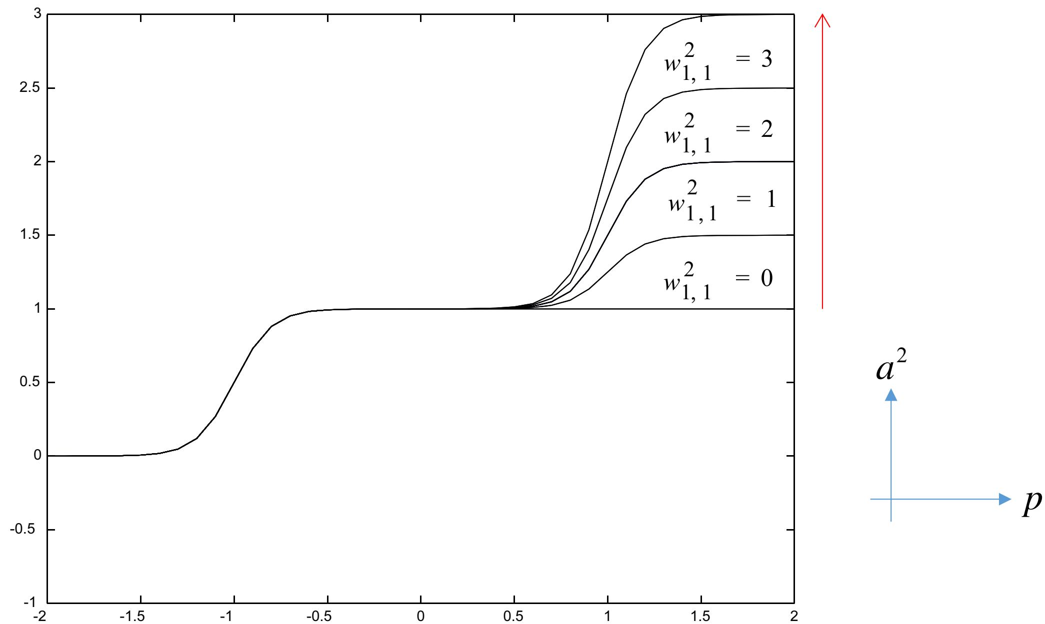


رگولاریزاسیون

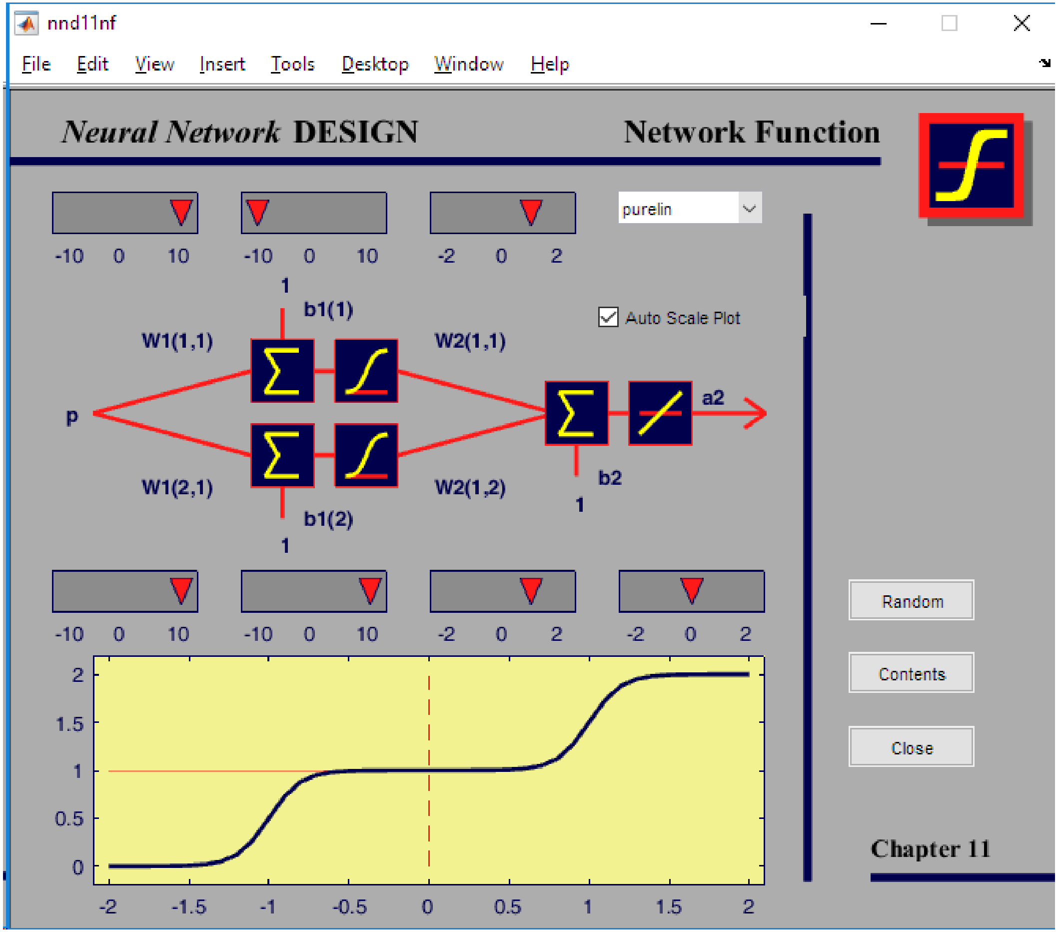
اثر تغییرات وزن

Effect of Weight Changes

افزایش یک وزن، شیب تابع شبکه را افزایش می‌دهد ⇐ احتمال بیش‌برازش به داده‌های آموزشی بیشتر می‌شود.



وقتی وزن‌ها محدودیت کوچک بودن داشته باشند، تابع شبکه یک درون‌یابی هموار از داده‌های آموزشی انجام می‌دهد:
(درست مانند زمانی که شبکه تعداد کوچکی نرون داشته باشد.)

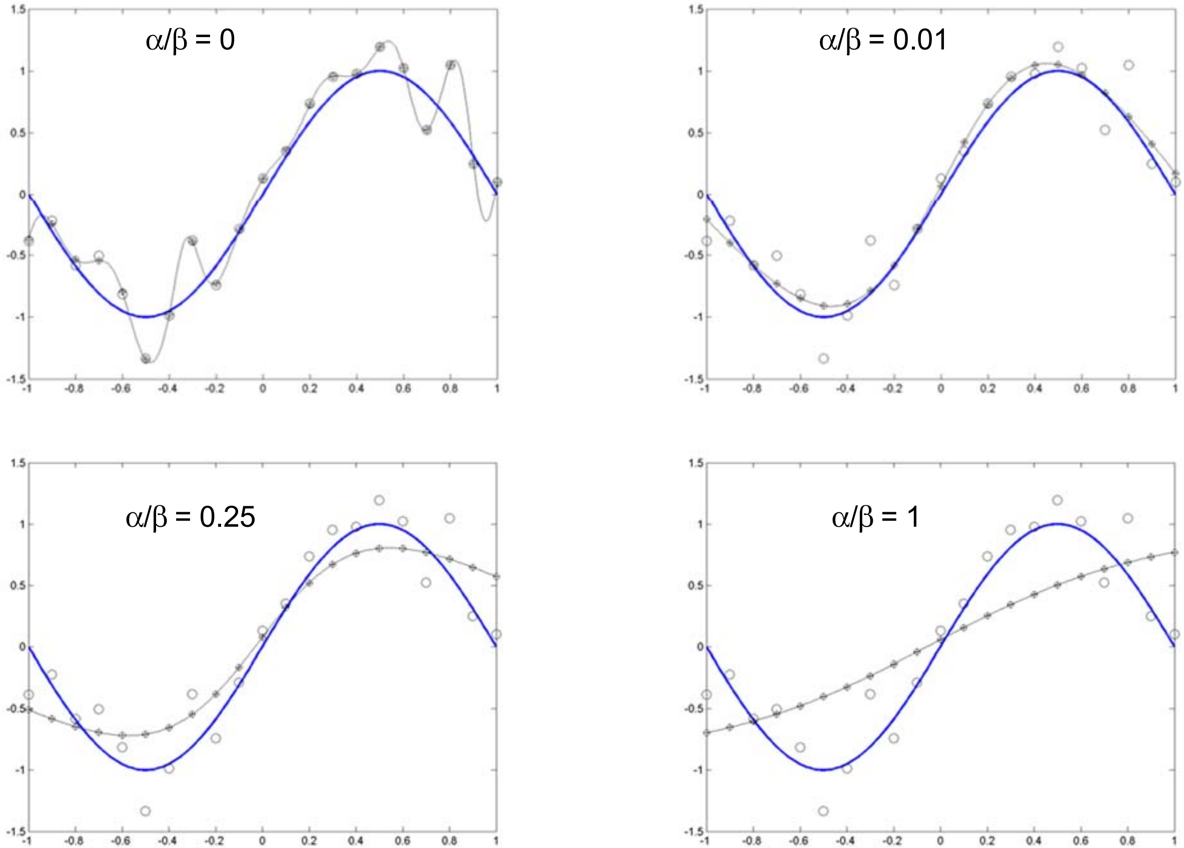


MATLAB

```
>> nnd11nf
```

13

Effect of Regularization

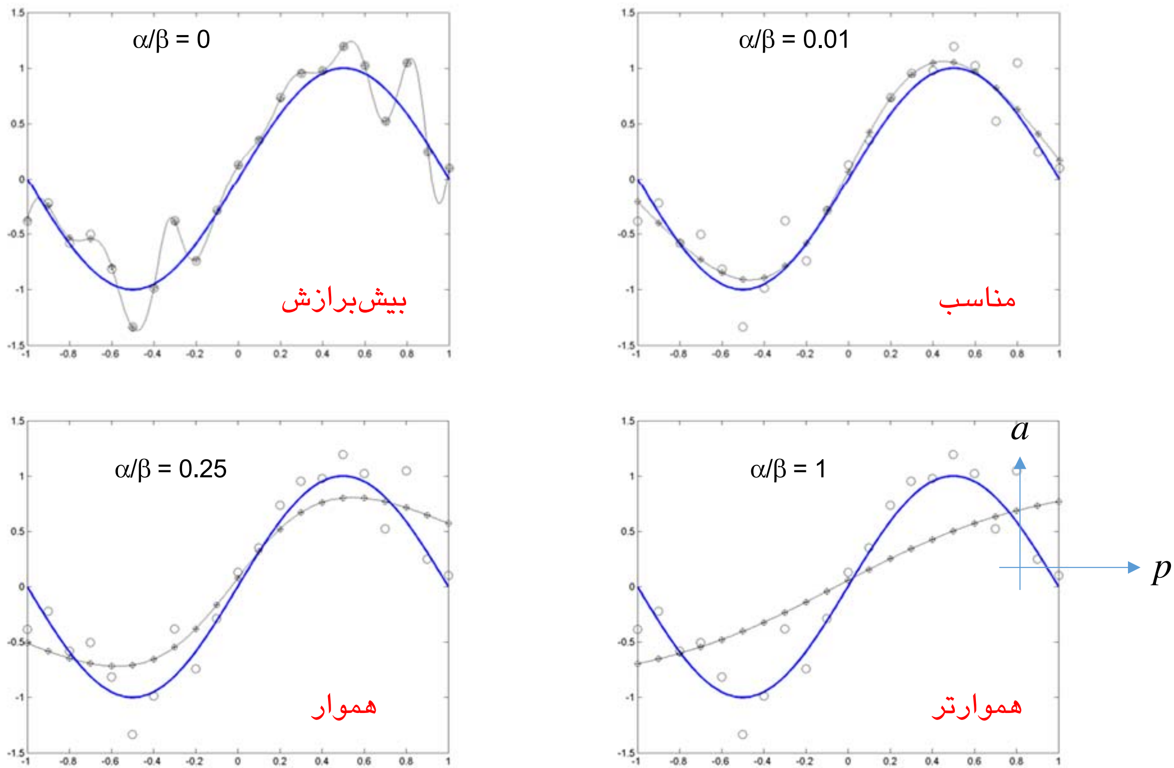


رگولاریزاسیون

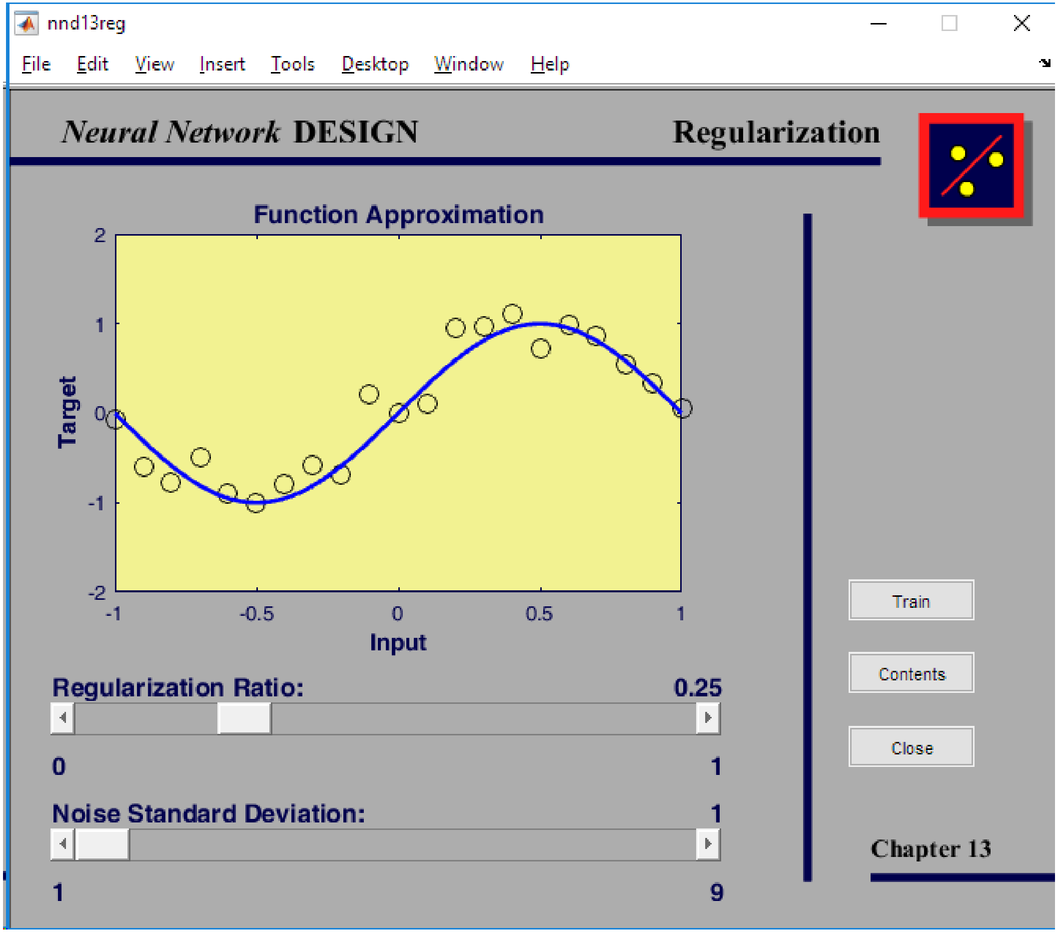
اثر رگولاریزاسیون

Effect of Regularization

یک شبکه‌ی 1-20-1 با ۲۱ نمونه‌ی نویزی از تابع سینوس



کلید موفقیت روش رگولاریزاسیون در تولید شبکه‌ی تعمیم‌پذیر، انتخاب مناسب نسبت α/β است.



MATLAB

```
>> nnd13reg
```

رگولاریزاسیون

تکنیک‌های تنظیم پارامترهای رگولاریزاسیون

REGULARIZATION

تکنیک‌های تنظیم پارامترهای رگولاریزاسیون	
	○ استفاده از مجموعه‌ی اعتبارسنجی (برای می‌نیمم کردن خطای مجموع مربعات روی آن) ○ روی‌کرد بیزی

شبکه‌های عصبی مصنوعی

تعمیم

۶

تحلیل بیزی

13

Bayes' Rule

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

$P(A)$ – Prior Probability. What we know about A before B is known.

$P(A|B)$ – Posterior Probability. What we know about A after we know the outcome of B.

$P(B|A)$ – Conditional Probability (Likelihood Function). Describes our knowledge of the system.

$P(B)$ – Marginal Probability. A normalization factor.

قاعده‌ی بیز

BAYES' RULE

$$P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)}$$

$P(A)$ – Prior Probability. What we know about A before B is known.

$P(A|B)$ – Posterior Probability. What we know about A after we know the outcome of B.

$P(B|A)$ – Conditional Probability (Likelihood Function). Describes our knowledge of the system.

$P(B)$ – Marginal Probability. A normalization factor.

Example Problem

- 1% of the population have a certain disease.
- A test for the disease is 80% accurate in detecting the disease in people who have it.
- 10% of the time the test yields a false positive.
- If you have a positive test, what is your probability of having the disease?

قاعده‌ی بیز

مثال (۱ از ۲)

BAYES' RULE

- 1% of the population have a certain disease.
- A test for the disease is 80% accurate in detecting the disease in people who have it.
- 10% of the time the test yields a false positive.
- If you have a positive test, what is your probability of having the disease?

Bayesian Analysis

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

A – Event that you have the disease.

B – Event that you have a positive test.

$P(A) = 0.01$

$P(B|A) = 0.8$

$P(B) = P(B|A)P(A) + P(B|\sim A)P(\sim A) = 0.8\ 0.01 + 0.1\ 0.99 = 0.107$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.8 \times 0.01}{0.107} = 0.0748$$

قاعده‌ی بیز

مثال (۲ از ۲)

BAYES' RULE

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

A – Event that you have the disease.

B – Event that you have a positive test.

$$P(A) = 0.01$$

$$P(B|A) = 0.8$$

$$P(B) = P(B|A)P(A) + P(B|\sim A)P(\sim A) = 0.8\ 0.01 + 0.1\ 0.99 = 0.107$$

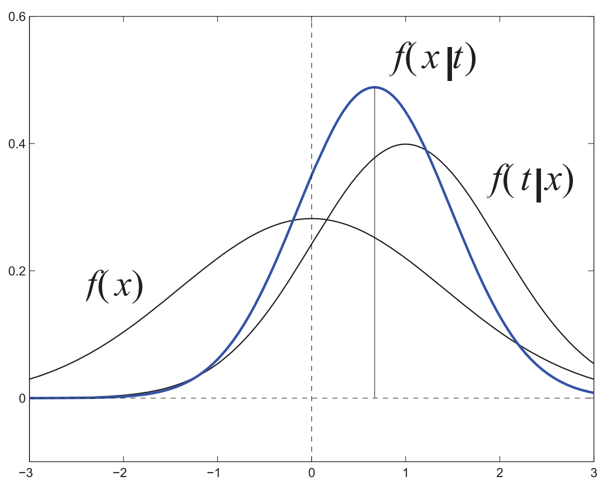
$$P(A\,|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.8 \times 0.01}{0.107} = 0.0748$$

Signal Plus Noise Example

$$t = x + \varepsilon$$

$$f(\varepsilon) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right) \qquad f(x) = \frac{1}{\sqrt{2\pi}\sigma_x} exp\left(-\frac{x^2}{2\sigma_x^2}\right)$$

$$f(t|x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(t-x)^2}{2\sigma^2}\right) \qquad f(x|t) = \frac{f(t|x)f(x)}{f(t)}$$



قاعده‌ی بیز

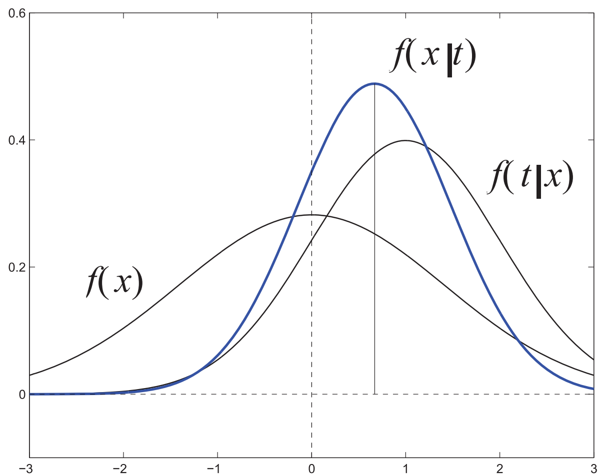
مثال: سیگنال به‌اضافه‌ی نویز

Signal Plus Noise Example

$$t = x + \varepsilon$$

$$f(\varepsilon) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right) \qquad f(x) = \frac{1}{\sqrt{2\pi}\sigma_x} exp\left(-\frac{x^2}{2\sigma_x^2}\right)$$

$$f(t|x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(t-x)^2}{2\sigma^2}\right) \qquad f(x|t) = \frac{f(t|x)f(x)}{f(t)}$$



تعمیم

۷

رگولاریزاسیون بیزی

تحلیل بیزی برای آموزش شبکه‌های عصبی چندلایه

در شبکه‌ی عصبی فرض پیشین هموار بودن تابع در حال تقریب وارد می‌شود

$\Downarrow$

وزن‌ها نمی‌توانند خیلی بزرگ باشند.

فرض: وزن‌های شبکه متغیر تصادفی هستند.

باید وزن‌هایی را انتخاب کنیم که احتمال شرطی وزن‌ها به‌شرط داده‌ها را ماکزیمم کند.

NN Bayesian Framework

(MacKay 92)

MP
Posterior

ML
Likelihood

Prior

$$P(\mathbf{x}\,|\,D, \alpha, \beta, M) = \frac{P(D\,|\,\mathbf{x}, \beta, M)\,P(\mathbf{x}\,|\,\alpha, M)}{P(D\,|\,\alpha, \beta, M)}$$

Normalization
(Evidence)

D - Data Set
M - Neural Network Model
$\mathbf{x}$ - Vector of Network Weights

تحلیل بیزی برای آموزش شبکه‌های عصبی چندلایه

چهارچوب بیزی برای شبکه‌ی عصبی

NN BAYESIAN FRAMEWORK

(MacKay 92)

MP
Posterior

ML
Likelihood

Prior

$$P(\mathbf{x} \mid D, \alpha, \beta, M) = \frac{P(D \mid \mathbf{x}, \beta, M)\, P(\mathbf{x} \mid \alpha, M)}{P(D \mid \alpha, \beta, M)}$$

بردار شامل
همه‌ی وزن‌ها و بایاس‌ها

Normalization
(Evidence)

D - Data Set

M - Neural Network Model (تعداد لایه‌ها / تعداد نرون در هر لایه)

$\mathbf{x}$ - Vector of Network Weights

α, β - Regularization Parameters

13

Gaussian Assumptions

Gaussian Noise

$$P(D|\mathbf{x},\beta,M) = \frac{1}{Z_D(\beta)}\exp(-\beta E_D) \qquad Z_D(\beta) = (2\pi\sigma_\varepsilon^2)^{N/2} = (\pi/\beta)^{N/2}$$

Gaussian Prior:

$$P(\mathbf{x}|\alpha,M) = \frac{1}{Z_W(\alpha)}\exp(-\alpha E_W) \qquad Z_W(\alpha) = (2\pi\sigma_w^2)^{n/2} = (\pi/\alpha)^{n/2}$$

$$P(\mathbf{x}|D,\alpha,\beta,M) = \frac{\frac{1}{Z_W(\alpha)}\frac{1}{Z_D(\beta)}\exp(-(\beta E_D + \alpha E_W))}{\text{Normalization Factor}} = \frac{1}{Z_F(\alpha,\beta)}\exp(-F(\mathbf{x}))$$

$$F = \beta E_D + \alpha E_W$$

Minimize F to Maximize P.

تحلیل بیزی برای آموزش شبکه‌های عصبی چندلایه

چهارچوب بیزی برای شبکه‌ی عصبی : فرض‌های گاوسی

GAUSSIAN ASSUMPTIONS

Gaussian Noise

$$\boxed{\beta = \frac{1}{2\sigma_\varepsilon^2}}$$

$$P(D|\mathbf{x}, \beta, M) = \frac{1}{Z_D(\beta)} \exp(-\beta E_D) \qquad Z_D(\beta) = (2\pi\sigma_\varepsilon^2)^{N/2} = (\pi/\beta)^{N/2}$$

Gaussian Prior:

$$\boxed{N = Q \times S^M}$$

$$P(\mathbf{x}|\alpha, M) = \frac{1}{Z_W(\alpha)} \exp(-\alpha E_W) \qquad Z_W(\alpha) = (2\pi\sigma_w^2)^{n/2} = (\pi/\alpha)^{n/2}$$

$$\boxed{\alpha = \frac{1}{2\sigma_w^2}}$$

$$P(\mathbf{x}|D, \alpha, \beta, M) = \frac{\frac{1}{Z_W(\alpha)} \frac{1}{Z_D(\beta)} \exp(-(\beta E_D + \alpha E_W))}{\text{Normalization Factor}} = \frac{1}{Z_F(\alpha, \beta)} \exp(-F(\mathbf{x}))$$

$$F = \beta E_D + \alpha E_W$$

Minimize F to Maximize P.

شاخص کارآیی: با آمار بیزی با فرض نویز گاوسی در مجموعه‌ی آموزشی و توزیع پیشین گاوسی برای وزن‌های شبکه.

* **هدف:** یافتن وزن‌ها برای ماکزیمم کردن توزیع پسین : محتمل‌ترین $\mathbf{x}^{MP}$

(در مقابل وزن‌هایی که تابع درست‌نمایی را ماکزیمم می‌کند : $\mathbf{x}^{ML}$)

تحلیل بیزی برای آموزش شبکه‌های عصبی چندلایه

چهارچوب بیزی برای شبکه‌ی عصبی: پارامترهای رگولاریزاسیون

REGULARIZATION PARAMETERS

معنای فیزیکی پارامترهای α و β بر اساس چهارچوب بیزی:

$$\beta = \frac{1}{2\sigma_\varepsilon^2}$$

با معکوس واریانس در نویز اندازه‌گیری ε_q متناسب است:

- اگر واریانس نویز بالا باشد $\Leftarrow$ β کوچک خواهد بود $\Leftarrow$ α/β بزرگ خواهد بود $\Leftarrow$ وزن‌های حاصل مجبور می‌شوند کوچک باشند و تابع شبکه هموار (smooth) می‌شود.
- هرچه نویز بزرگ‌تر باشد $\Leftarrow$ تابع شبکه را هموارتر می‌کنیم: با متوسطگیری از اثرات نویز

$$\alpha = \frac{1}{2\sigma_w^2}$$

با معکوس واریانس توزیع پیشین وزن‌های شبکه متناسب است:

- اگر این واریانس بالا باشد $\Leftarrow$ اطمینان کمی در مورد مقدار وزن‌های شبکه داریم $\Leftarrow$ وزن‌ها می‌توانند خیلی بزرگ باشند $\Leftarrow$ α کوچک خواهد بود $\Leftarrow$ α/β کوچک خواهد بود $\Leftarrow$ بزرگ بودن وزن‌های شبکه و تغییرات بیشتر در تابع شبکه
- هرچه این واریانس بزرگ‌تر باشد $\Leftarrow$ تغییرات بیشتری برای تابع شبکه مجاز می‌شود.

13 Optimizing Regularization Parameters

Second Level of Inference

$$P(\alpha, \beta | D, M) = \frac{\overbrace{P(D | \alpha, \beta, M)}^{\text{Evidence from First Level}} P(\alpha, \beta | M)}{P(D | M)}$$

Evidence:

$$P(D | \alpha, \beta, M) = \frac{P(D | \mathbf{x}, \beta, M) P(\mathbf{x} | \alpha, M)}{P(\mathbf{x} | D, \alpha, \beta, M)}$$

$$= \frac{\left[\frac{1}{Z_D(\beta)} \exp(-\beta E_D)\right]\left[\frac{1}{Z_W(\alpha)} \exp(-\alpha E_W)\right]}{\frac{1}{Z_F(\alpha, \beta)} \exp(-F(\mathbf{x}))}$$

$$= \frac{Z_F(\alpha, \beta)}{Z_D(\beta) Z_W(\alpha)} \cdot \frac{\exp(-\beta E_D - \alpha E_W)}{\exp(-F(\mathbf{x}))} = \frac{Z_F(\alpha, \beta)}{Z_D(\beta) Z_W(\alpha)}$$

$Z_F(\alpha, \beta)$ is the only unknown in this expression.

تحلیل بیزی برای آموزش شبکه‌های عصبی چندلایه

چهارچوب بیزی برای شبکه‌ی عصبی: بهینه‌سازی پارامترهای رگولاریزاسیون

OPTIMIZING REGULARIZATION PARAMETERS

هدف: یافتن راهی برای تخمین پارامترهای α و β از روی داده‌ها

$\Leftarrow$ نیاز به تحلیل بیزی در یک سطح بالاتر

Second Level of Inference

$$P(\alpha, \beta | D, M) = \frac{\overbrace{P(D|\alpha, \beta, M)}^{\text{Evidence from First Level}} P(\alpha, \beta | M)}{P(D|M)}$$

Evidence:

$$P(D|\alpha, \beta, M) = \frac{P(D|\mathbf{x}, \beta, M) P(\mathbf{x}|\alpha, M)}{P(\mathbf{x}|D, \alpha, \beta, M)}$$

با فرض گاوسی بودن همه‌ی احتمالات

$$= \frac{\left[\frac{1}{Z_D(\beta)} \exp(-\beta E_D)\right]\left[\frac{1}{Z_W(\alpha)} \exp(-\alpha E_W)\right]}{\frac{1}{Z_F(\alpha, \beta)} \exp(-F(\mathbf{x}))}$$

$$= \frac{Z_F(\alpha, \beta)}{Z_D(\beta) Z_W(\alpha)} \cdot \frac{\exp(-\beta E_D - \alpha E_W)}{\exp(-F(\mathbf{x}))} = \frac{Z_F(\alpha, \beta)}{Z_D(\beta) Z_W(\alpha)}$$

$Z_F(\alpha, \beta)$ is the only unknown in this expression.

Z_F را با بسط تیلور تقریب می‌زنیم.

Quadratic Approximation

Taylor series expansion:

$$F(\mathbf{x}) \approx F(\mathbf{x}^{MP}) + \frac{1}{2}(\mathbf{x} - \mathbf{x}^{MP})^T \mathbf{H}^{MP}(\mathbf{x} - \mathbf{x}^{MP}) \qquad \mathbf{H} = \beta \nabla^2 E_D + \alpha \nabla^2 E_W$$

Substituting into previous posterior density function:

$$P(\mathbf{x} | D, \alpha, \beta, M) \approx \frac{1}{Z_F} \exp\left[-F(\mathbf{x}^{MP}) - \frac{1}{2}(\mathbf{x} - \mathbf{x}^{MP})^T \mathbf{H}^{MP}(\mathbf{x} - \mathbf{x}^{MP})\right]$$

$$P(\mathbf{x} | D, \alpha, \beta, M) \approx \left\{\frac{1}{Z_F} \exp(-F(\mathbf{x}^{MP}))\right\} \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{x}^{MP})^T \mathbf{H}^{MP}(\mathbf{x} - \mathbf{x}^{MP})\right]$$

Equate with standard Gaussian density:

$$P(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \left|(\mathbf{H}^{MP})^{-1}\right|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}^{MP})^T \mathbf{H}^{MP}(\mathbf{x} - \mathbf{x}^{MP})\right)$$

Comparing to previous equation, we have:

$$Z_F(\alpha, \beta) \approx (2\pi)^{n/2} (\det((\mathbf{H}^{MP})^{-1}))^{1/2} \exp(-F(\mathbf{x}^{MP}))$$

تحلیل بیزی برای آموزش شبکه‌های عصبی چندلایه

چهارچوب بیزی برای شبکه‌ی عصبی: بهینه‌سازی پارامترهای رگولاریزاسیون: تقریب درجه دوم

QUADRATIC APPROXIMATION

Taylor series expansion:

بسط سری تیلور حول می‌نیمم تابع F شکل درجه دوم و گرادیان صفر

$$F(\mathbf{x}) \approx F(\mathbf{x}^{MP}) + \frac{1}{2}(\mathbf{x} - \mathbf{x}^{MP})^T \mathbf{H}^{MP} (\mathbf{x} - \mathbf{x}^{MP}) \qquad \mathbf{H} = \beta \nabla^2 E_D + \alpha \nabla^2 E_W$$

Substituting into previous posterior density function:

جایگذاری:

$$P(\mathbf{x} | D, \alpha, \beta, M) \approx \frac{1}{Z_F} \exp\left[-F(\mathbf{x}^{MP}) - \frac{1}{2}(\mathbf{x} - \mathbf{x}^{MP})^T \mathbf{H}^{MP} (\mathbf{x} - \mathbf{x}^{MP})\right]$$

$$P(\mathbf{x} | D, \alpha, \beta, M) \approx \left\{\frac{1}{Z_F} \exp(-F(\mathbf{x}^{MP}))\right\} \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{x}^{MP})^T \mathbf{H}^{MP} (\mathbf{x} - \mathbf{x}^{MP})\right]$$

Equate with standard Gaussian density:

$$P(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \left|(\mathbf{H}^{MP})^{-1}\right|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}^{MP})^T \mathbf{H}^{MP} (\mathbf{x} - \mathbf{x}^{MP})\right)$$

Comparing to previous equation, we have:

$$Z_F(\alpha, \beta) \approx (2\pi)^{n/2} (\det((\mathbf{H}^{MP})^{-1}))^{1/2} \exp(-F(\mathbf{x}^{MP}))$$

Optimum Parameters

If we make this substitution for Z_F in the expression for the evidence and then take the derivative with respect to α and β to locate the minimum we find:

$$\alpha^{MP} = \frac{\gamma}{2E_W(\mathbf{x}^{MP})} \qquad \beta^{MP} = \frac{N-\gamma}{2E_D(\mathbf{x}^{MP})}$$

Effective Number of Parameters

$$\gamma = n - 2\alpha^{MP}\mathrm{tr}(\mathbf{H}^{MP})^{-1}$$

تحلیل بیزی برای آموزش شبکه‌های عصبی چندلایه

Optimum Parameters

If we make this substitution for Z_F in the expression for the evidence and then take the derivative with respect to α and β to locate the minimum we find:

$$\alpha^{MP} = \frac{\gamma}{2E_W(\mathbf{x}^{MP})} \qquad \beta^{MP} = \frac{N-\gamma}{2E_D(\mathbf{x}^{MP})}$$

تعداد مؤثر پارامترها

Effective Number of Parameters

چه تعدادی از پارامترهای شبکه (وزن / بایاس) در کاهش دادن تابع خطا به‌طور مؤثر استفاده می‌شوند؟

$$\gamma = n - 2\alpha^{MP}\mathrm{tr}(\mathbf{H}^{MP})^{-1}$$

n = تعداد کل پارامترهای شبکه $\qquad 0 \le \gamma \le n$

13

Gauss-Newton Approximation

It can be expensive to compute the Hessian matrix.

Try the Gauss-Newton Approximation.

$$\mathbf{H} = \nabla^2 F(\mathbf{x}) \approx 2\beta \mathbf{J}^T\mathbf{J} + 2\alpha \mathbf{I}_n$$

This is readily available if the Levenberg-Marquardt algorithm is used for training.

تحلیل بیزی برای آموزش شبکه‌های عصبی چندلایه

چهارچوب بیزی برای شبکه‌ی عصبی: بهینه‌سازی پارامترهای رگولاریزاسیون: تقریب گاوس–نیوتن

GAUSS-NEWTON APPROXIMATION

It can be expensive to compute the Hessian matrix.

Try the Gauss-Newton Approximation.

$$\mathbf{H} = \nabla^2 F(\mathbf{x}) \approx 2\beta \mathbf{J}^T\mathbf{J} + 2\alpha \mathbf{I}_n$$

تقریب برای ماتریس هسی

This is readily available if the Levenberg-Marquardt algorithm is used for training.

Algorithm (GNBR)

1. Initialize α, β and the weights.

2. Take one step of Levenberg-Marquardt to minimize $F(\mathbf{w})$.

3. Compute the effective number of parameters $\gamma = n - 2\alpha\,\mathrm{tr}(\mathbf{H}^{-1})$, using the Gauss-Newton approximation for $\mathbf{H}$.

4. Compute new estimates of the regularization parameters $\alpha = \gamma/(2E_W)$ and $\beta = (N - \gamma)/(2E_D)$.

5. Iterate steps 1-3 until convergence.

تحلیل بیزی برای آموزش شبکه‌های عصبی چندلایه

چهارچوب بیزی برای شبکه‌ی عصبی : بهینه‌سازی پارامترهای رگولاریزاسیون : تقریب گاوس–نیوتن : الگوریتم

GAUSS-NEWTON BAYESIAN REGULARIZATION ALGORITHM (GNBR)

چهارچوب بیزی برای رگولاریزاسیون با استفاده از تقریب ماتریس هسی با روش گاوس–نیوتن :

وزن‌های تصادفی

1. Initialize α, β and the weights.

محاسبه‌ی F

2. Take one step of Levenberg-Marquardt to minimize $F(\mathbf{w})$.

$$\boxed{F(\mathbf{w}) = \beta E_D + \alpha E_W}$$

انتخاب γ

$\boxed{0 \le \gamma \le n}$

3. Compute the effective number of parameters
$\gamma = n - 2\alpha\, \mathrm{tr}(\mathbf{H}^{-1})$,
using the Gauss-Newton approximation for $\mathbf{H}$.

محاسبه‌ی α و β

4. Compute new estimates of the regularization parameters
$\alpha = \gamma/(2E_W)$ and $\beta = (N - \gamma)/(2E_D)$.

تکرار تا همگرایی

5. Iterate steps 1-3 until convergence.

با هر بار تخمین α و β تابع $F(\mathbf{w})$ عوض می‌شود $\Leftarrow$ نقطه‌ی می‌نیمم تغییر می‌کند.

بهترین نتایج GNBR زمانی حاصل می‌شود که داده‌های ورودی به بازه‌ی $[-1,1]$ نگاشت پیدا کنند.

Checks of Performance

- If γ is very close to n, then the network may be too small.
  - Add more hidden layer neurons and retrain.
- If the larger network has the same final γ, then the smaller network was large enough.
- Otherwise, increase the number of hidden neurons.
- If a network is sufficiently large, then a larger network will achieve comparable values for γ, E_D and E_W.

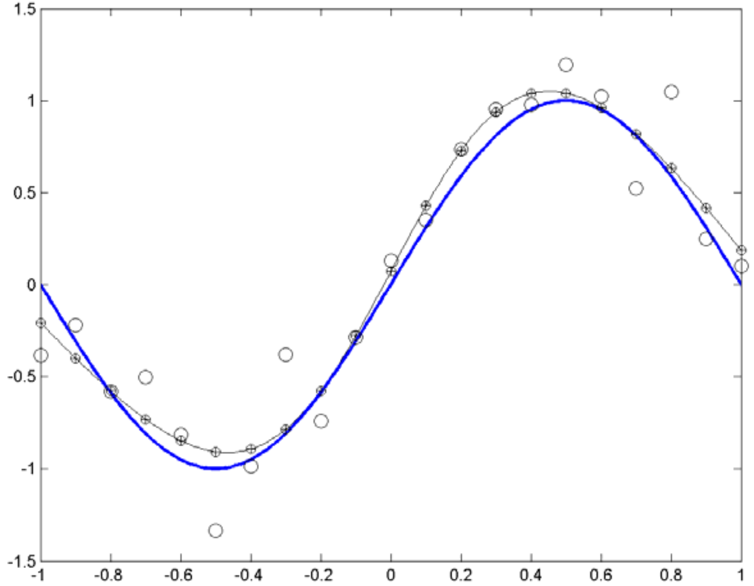
تحلیل بیزی برای آموزش شبکه‌های عصبی چندلایه

CHECKS OF PERFORMANCE

- If γ is very close to n, then the network may be too small.
- Add more hidden layer neurons and retrain.
- If the larger network has the same final γ, then the smaller network was large enough.
- Otherwise, increase the number of hidden neurons.
- If a network is sufficiently large, then a larger network will achieve comparable values for γ, E_D and E_W.

GNBR Example

$\alpha/\beta = 0.0137$

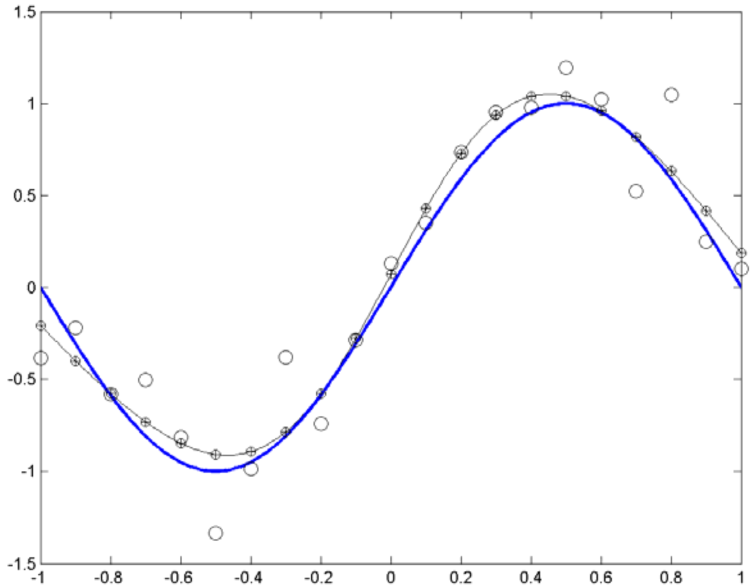


تحلیل بیزی برای آموزش شبکه‌های عصبی چندلایه

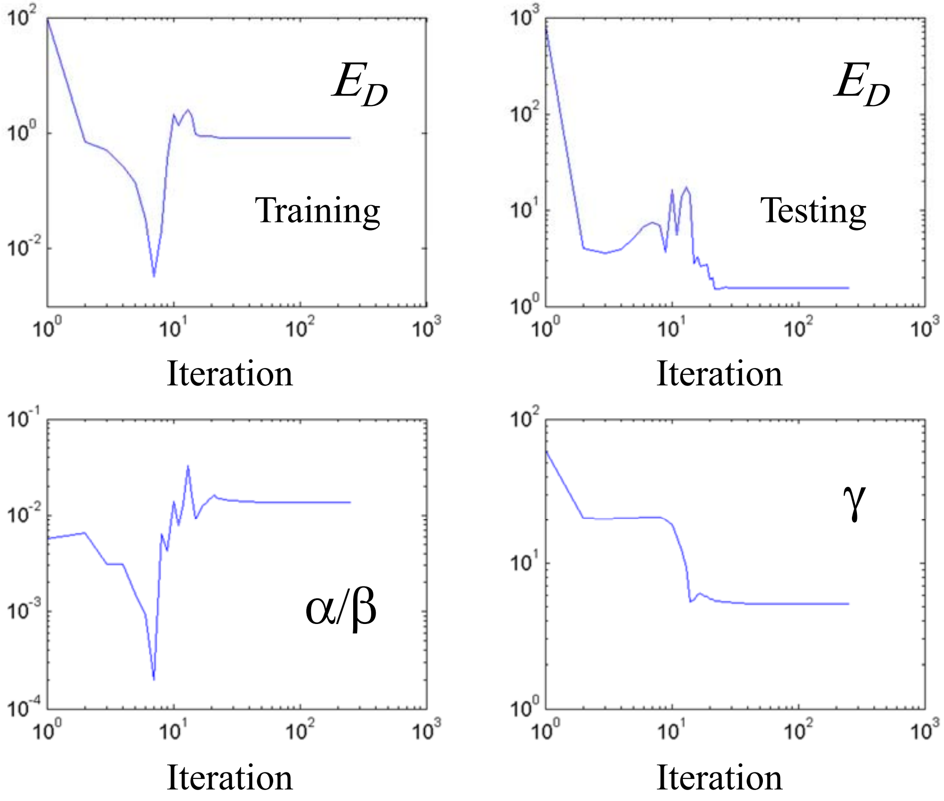
GNBR EXAMPLE

$$\alpha/\beta = 0.0137$$

شبکه‌ی 1-20-1
کلاً ۶۱ وزن



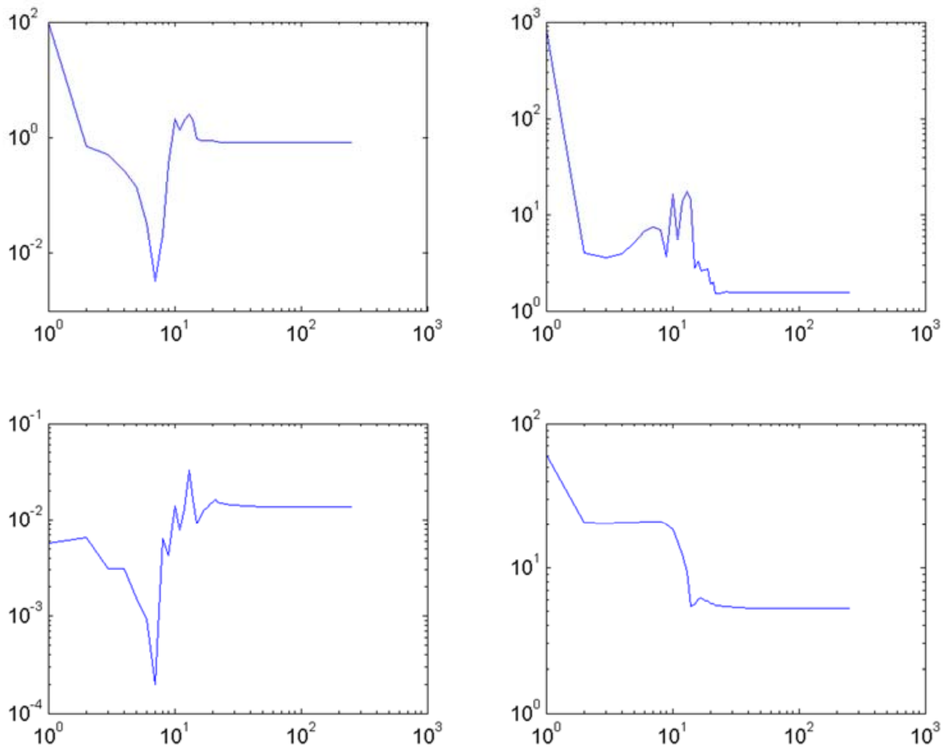
Convergence of GNBR



تحلیل بیزی برای آموزش شبکه‌های عصبی چندلایه

چهارچوب بیزی برای شبکه‌ی عصبی: بهینه‌سازی پارامترهای رگولاریزاسیون: تقریب گاوس–نیوتن: الگوریتم: همگرایی

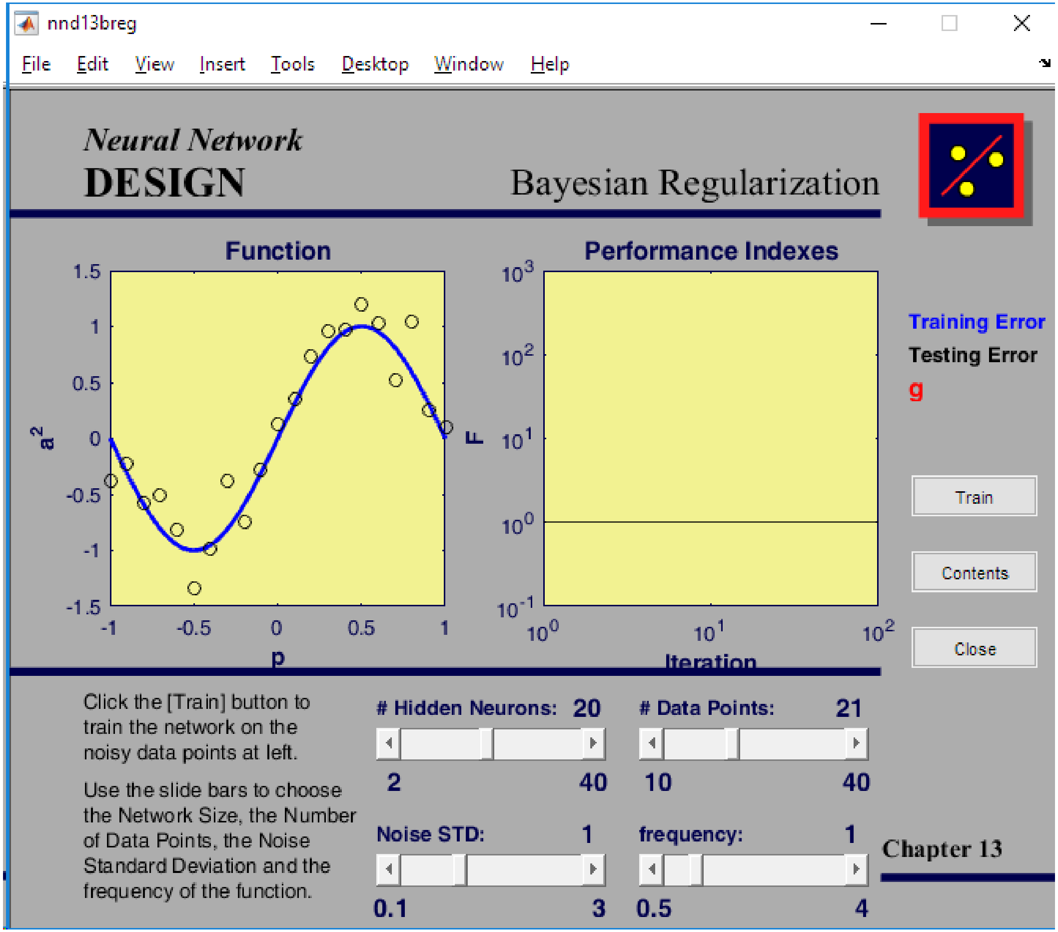
Convergence of GNBR



شبکه‌ی 1-20-1
کلاً ۶۱ وزن

γ نهایی = 5.2
(یعنی ۶ وزن از ۶۱ وزن)

معایب شبکه‌ی بزرگ‌تر: بیش‌برازش + محاسبات بیشتر


nnd13breg
File Edit View Insert Tools Desktop Window Help
Neural Network
DESIGN
Bayesian Regularization
Function
Performance Indexes
Training Error
Testing Error
g
Train
Contents
Close
p
Iteration
Click the [Train] button to train the network on the noisy data points at left.
Use the slide bars to choose the Network Size, the Number of Data Points, the Noise Standard Deviation and the frequency of the function.
Hidden Neurons: 20
2 40
Data Points: 21
10 40
Noise STD: 1
0.1 3
frequency: 1
0.5 4
Chapter 13

>> nnd13breg

تعمیم

۸

رابطه‌ی بین رگولاریزاسیون و توقف زودهنگام

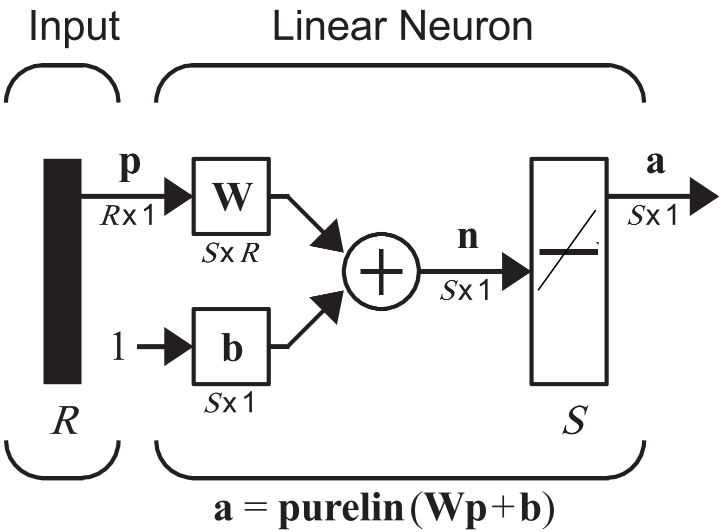
Relationship between Early Stopping and Regularization

رابطه‌ی بین رگولاریزاسیون و توقف زودهنگام

RELATIONSHIP BETWEEN EARLY STOPPING AND REGULARIZATION

نشان دادن هم‌ارزی تقریبی بین دو روش افزایش تعمیم‌پذیری
با استفاده از یک مثال خطی

Linear Network



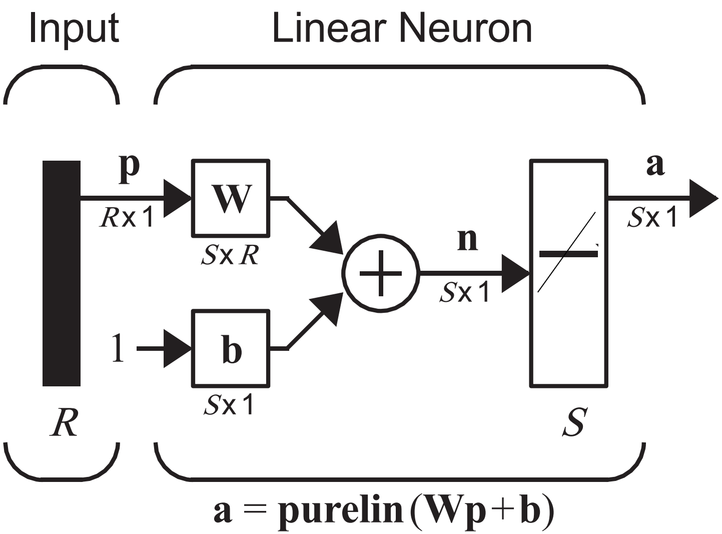
$$\mathbf{a} = \mathbf{purelin}(\mathbf{W}\mathbf{p} + \mathbf{b}) = \mathbf{W}\mathbf{p} + \mathbf{b}$$

$$a_i = purelin(n_i) = purelin({}_i\mathbf{w}^T\mathbf{p} + b_i) = {}_i\mathbf{w}^T\mathbf{p} + b_i \qquad {}_i\mathbf{w} = \begin{bmatrix} w_{i,1} \\ w_{i,2} \\ \vdots \\ w_{i,R} \end{bmatrix}$$

رابطه‌ی بین رگولاریزاسیون و توقف زودهنگام

شبکه‌ی خطی

Linear Network



$$\mathbf{a} = \mathrm{purelin}(\mathbf{Wp} + \mathbf{b}) = \mathbf{Wp} + \mathbf{b}$$

$$a_i = purelin(n_i) = purelin({}_i\mathbf{w}^T\mathbf{p} + b_i) = {}_i\mathbf{w}^T\mathbf{p} + b_i$$

$${}_i\mathbf{w} = \begin{bmatrix} w_{i,1} \\ w_{i,2} \\ \vdots \\ w_{i,R} \end{bmatrix}$$

Performance Index

Training Set:

$$\{\mathbf{p}_1, \mathbf{t}_1\}, \{\mathbf{p}_2, \mathbf{t}_2\}, \dots, \{\mathbf{p}_Q, \mathbf{t}_Q\}$$

Input: $\mathbf{p}_q$ Target: $\mathbf{t}_q$

Notation:

$$\mathbf{x} = \begin{bmatrix} {}_1\mathbf{w} \\ b \end{bmatrix} \qquad \mathbf{z} = \begin{bmatrix} \mathbf{p} \\ 1 \end{bmatrix} \qquad a = {}_1\mathbf{w}^T \mathbf{p} + b \quad \Rightarrow \quad a = \mathbf{x}^T \mathbf{z}$$

Mean Square Error:

$$F(\mathbf{x}) = E[e^2] = E[(t-a)^2] = E[(t - \mathbf{x}^T\mathbf{z})^2] = E_D$$

رابطه‌ی بین رگولاریزاسیون و توقف زودهنگام

شاخص کارآیی

PERFORMANCE INDEX

Training Set:

$$\{\mathbf{p}_1, \mathbf{t}_1\}, \{\mathbf{p}_2, \mathbf{t}_2\}, \ldots, \{\mathbf{p}_Q, \mathbf{t}_Q\}$$

Input: $\mathbf{p}_q$ Target: $\mathbf{t}_q$

Notation:

$$\mathbf{x} = \begin{bmatrix} {}_1\mathbf{w} \\ b \end{bmatrix} \qquad \mathbf{z} = \begin{bmatrix} \mathbf{p} \\ 1 \end{bmatrix} \qquad a = {}_1\mathbf{w}^T\mathbf{p} + b \Rightarrow a = \mathbf{x}^T\mathbf{z}$$

Mean Square Error:

$$F(\mathbf{x}) = E[e^2] = E[(t-a)^2] = E[(t-\mathbf{x}^T\mathbf{z})^2] = E_D$$

Error Analysis

$$F(\mathbf{x}) = E[e^2] = E[(t-a)^2] = E[(t-\mathbf{x}^T\mathbf{z})^2]$$

$$F(\mathbf{x}) = E[t^2 - 2t\mathbf{x}^T\mathbf{z} + \mathbf{x}^T\mathbf{z}\mathbf{z}^T\mathbf{x}]$$

$$F(\mathbf{x}) = E[t^2] - 2\mathbf{x}^T \mathrm{E}[t\mathbf{z}] + \mathbf{x}^T E[\mathbf{z}\mathbf{z}^T]\mathbf{x}$$

$$\boxed{F(\mathbf{x}) = c - 2\mathbf{x}^T\mathbf{h} + \mathbf{x}^T\mathbf{R}\mathbf{x}}$$

$$c = E[t^2] \qquad \mathbf{h} = E[t\mathbf{z}] \qquad \mathbf{R} = E[\mathbf{z}\mathbf{z}^T]$$

The mean square error for the Linear Network is a quadratic function:

$$F(\mathbf{x}) = c + \mathbf{d}^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}$$

$$\mathbf{d} = -2\mathbf{h} \qquad \mathbf{A} = 2\mathbf{R}$$

رابطه‌ی بین رگولاریزاسیون و توقف زودهنگام

تحلیل خطا

Error Analysis

$$F(\mathbf{x}) = E[e^2] = E[(t-a)^2] = E[(t-\mathbf{x}^T\mathbf{z})^2]$$

$$F(\mathbf{x}) = E[t^2 - 2t\mathbf{x}^T\mathbf{z} + \mathbf{x}^T\mathbf{z}\mathbf{z}^T\mathbf{x}]$$

$$F(\mathbf{x}) = E[t^2] - 2\mathbf{x}^T E[t\mathbf{z}] + \mathbf{x}^T E[\mathbf{z}\mathbf{z}^T]\mathbf{x}$$

$$\boxed{F(\mathbf{x}) = c - 2\mathbf{x}^T\mathbf{h} + \mathbf{x}^T\mathbf{R}\mathbf{x}}$$

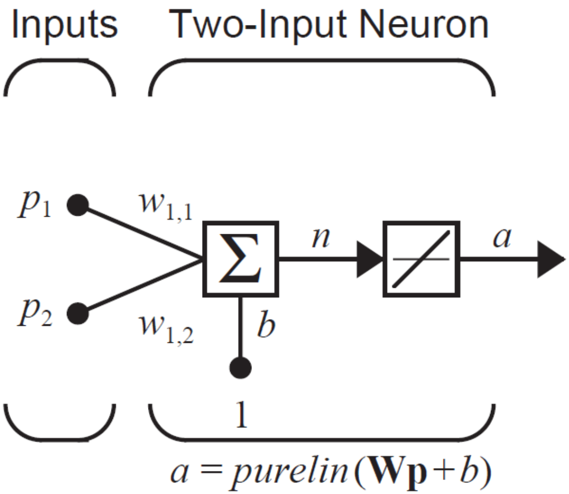
$$c = E[t^2] \qquad \mathbf{h} = E[t\mathbf{z}] \qquad \mathbf{R} = E[\mathbf{z}\mathbf{z}^T]$$

The mean square error for the Linear Network is a quadratic function:

$$F(\mathbf{x}) = c + \mathbf{d}^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}$$

$$\mathbf{d} = -2\mathbf{h} \qquad \mathbf{A} = 2\mathbf{R}$$

Example



$$\left\{\mathbf{p}_1 = \begin{bmatrix}1\\1\end{bmatrix}, t_1 = 1\right\}$$ (Probability = 0.75)

$$\left\{\mathbf{p}_2 = \begin{bmatrix}-1\\1\end{bmatrix}, t_2 = -1\right\}$$ (Probability = 0.25)

$$F(\mathbf{x}) = c - 2\mathbf{x}^T\mathbf{h} + \mathbf{x}^T\mathbf{R}\mathbf{x} = E_D$$

$$c = E[t^2] = (1)^2(0.75) + (-1)^2(0.25) = 1$$

$$\mathbf{h} = E[t\mathbf{z}] = (0.75)(1)\begin{bmatrix}1\\1\end{bmatrix} + (0.25)(-1)\begin{bmatrix}-1\\1\end{bmatrix} = \begin{bmatrix}1\\0.5\end{bmatrix}$$

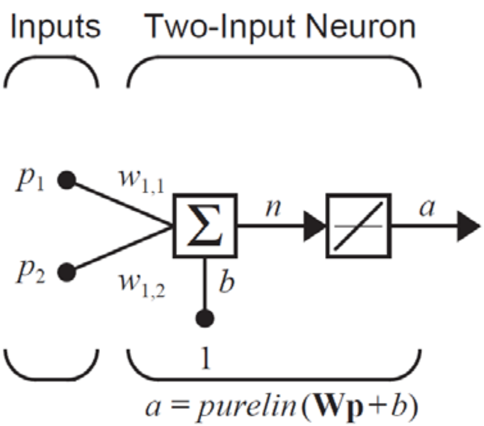
$$\mathbf{R} = E[\mathbf{z}\mathbf{z}^T] = \mathbf{p}_1\mathbf{p}_1^T(0.75) + \mathbf{p}_2\mathbf{p}_2^T(0.25)$$

$$= 0.75\begin{bmatrix}1\\1\end{bmatrix}\begin{bmatrix}1 & 1\end{bmatrix} + 0.25\begin{bmatrix}-1\\1\end{bmatrix}\begin{bmatrix}-1 & 1\end{bmatrix} = \begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}$$

رابطه‌ی بین رگولاریزاسیون و توقف زودهنگام

مثال

Example



$$\left\{\mathbf{p}_1 = \begin{bmatrix}1\\1\end{bmatrix},\ t_1 = 1\right\} \quad \text{(Probability = 0.75)}$$

$$\left\{\mathbf{p}_2 = \begin{bmatrix}-1\\1\end{bmatrix},\ t_2 = -1\right\} \quad \text{(Probability = 0.25)}$$

$$F(\mathbf{x}) = c - 2\mathbf{x}^T\mathbf{h} + \mathbf{x}^T\mathbf{R}\mathbf{x} = E_D$$

$$c = E[t^2] = (1)^2(0.75) + (-1)^2(0.25) = 1$$

$$\mathbf{h} = E[t\mathbf{z}] = (0.75)(1)\begin{bmatrix}1\\1\end{bmatrix} + (0.25)(-1)\begin{bmatrix}-1\\1\end{bmatrix} = \begin{bmatrix}1\\0.5\end{bmatrix}$$

$$\mathbf{R} = E[\mathbf{z}\mathbf{z}^T] = \mathbf{p}_1\mathbf{p}_1^T(0.75) + \mathbf{p}_2\mathbf{p}_2^T(0.25)$$

$$= 0.75\begin{bmatrix}1\\1\end{bmatrix}\begin{bmatrix}1 & 1\end{bmatrix} + 0.25\begin{bmatrix}-1\\1\end{bmatrix}\begin{bmatrix}-1 & 1\end{bmatrix} = \begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}$$

Performance Contour

Optimum Point (Maximum Likelihood)

$$\mathbf{x}^{ML} = -\mathbf{A}^{-1}\mathbf{d} = \mathbf{R}^{-1}\mathbf{h} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Hessian Matrix

$$\nabla^2 F(\mathbf{x}) = \mathbf{A} = 2\mathbf{R} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

Eigenvalues

$$\left| \mathbf{A} - \lambda \mathbf{I} \right| = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = \lambda^2 - 4\lambda + 3 = (\lambda - 1)(\lambda - 3) \Rightarrow \lambda_1 = 1, \qquad \lambda_2 = 3$$

Eigenvectors

$$[\mathbf{A} - \lambda \mathbf{I}]\mathbf{v} = 0$$

$$\lambda_1 = 1 \qquad \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \mathbf{v}_1 = 0 \qquad \mathbf{v}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \qquad \lambda_2 = 3 \qquad \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \mathbf{v}_2 = 0 \qquad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

رابطه‌ی بین رگولاریزاسیون و توقف زودهنگام

کانتور کارآیی

PERFORMANCE CONTOUR

Optimum Point (Maximum Likelihood)

$$\mathbf{x}^{ML} = -\mathbf{A}^{-1}\mathbf{d} = \mathbf{R}^{-1}\mathbf{h} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Hessian Matrix

$$\nabla^2 F(\mathbf{x}) = \mathbf{A} = 2\mathbf{R} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

Eigenvalues

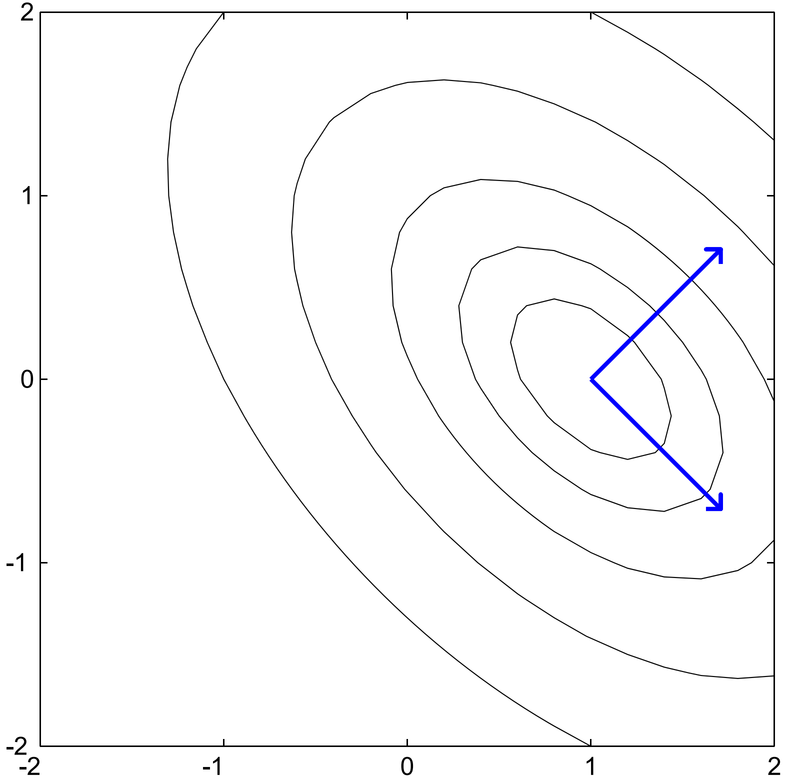
$$\left| \mathbf{A} - \lambda\mathbf{I} \right| = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = \lambda^2 - 4\lambda + 3 = (\lambda - 1)(\lambda - 3) \Rightarrow \lambda_1 = 1, \quad \lambda_2 = 3$$

Eigenvectors

$$[\mathbf{A} - \lambda\mathbf{I}]\mathbf{v} = 0$$

$$\lambda_1 = 1 \quad \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\mathbf{v}_1 = 0 \quad \mathbf{v}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \qquad \lambda_2 = 3 \quad \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}\mathbf{v}_2 = 0 \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

13 Contour Plot of E_D

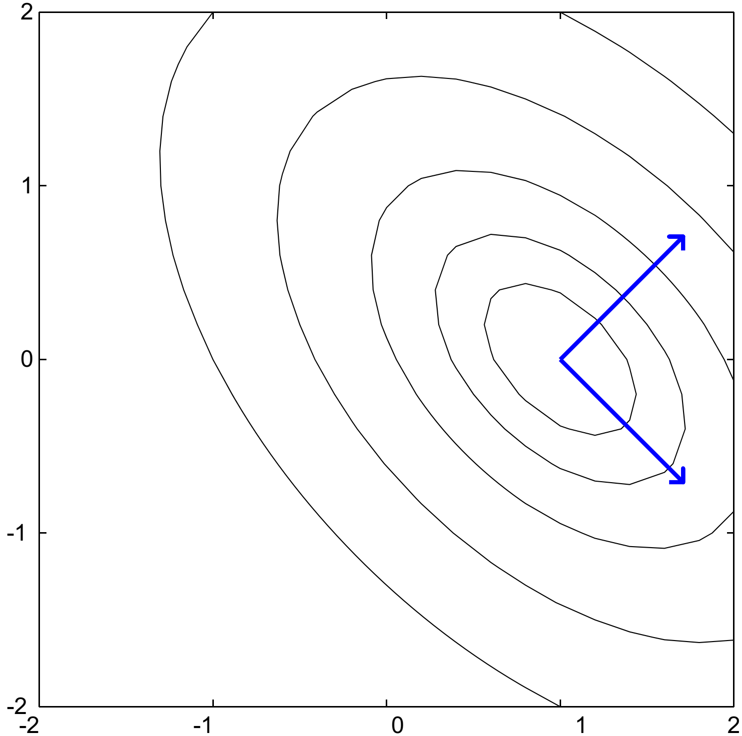


$$\gamma = n - 2\alpha^{MP}\mathrm{tr}(\mathbf{H}^{MP})^{-1}$$

رابطه‌ی بین رگولاریزاسیون و توقف زودهنگام

رسم کانتور E_D

CONTOUR PLOT OF E_D



$$\gamma = n - 2\alpha^{MP}\mathrm{tr}(\mathbf{H}^{MP})^{-1}$$

Steepest Descent Trajectory

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \mathbf{g}_k = \mathbf{x}_k - \alpha(\mathbf{A}\mathbf{x}_k + \mathbf{d})$$

$$= \mathbf{x}_k - \alpha \mathbf{A}(\mathbf{x}_k + \mathbf{A}^{-1}\mathbf{d}) = \mathbf{x}_k - \alpha \mathbf{A}(\mathbf{x}_k - \mathbf{x}^{ML})$$

$$= [\mathbf{I} - \alpha \mathbf{A}]\mathbf{x}_k + \alpha \mathbf{A}\mathbf{x}^{ML} = \mathbf{M}\mathbf{x}_k + [\mathbf{I} - \mathbf{M}]\mathbf{x}^{ML}$$

$$\boxed{\mathbf{M} = [\mathbf{I} - \alpha \mathbf{A}]}$$

$$\mathbf{x}_1 = \mathbf{M}\mathbf{x}_0 + [\mathbf{I} - \mathbf{M}]\mathbf{x}^{ML}$$

$$\mathbf{x}_2 = \mathbf{M}\mathbf{x}_1 + [\mathbf{I} - \mathbf{M}]\mathbf{x}^{ML}$$

$$= \mathbf{M}^2\mathbf{x}_0 + \mathbf{M}[\mathbf{I} - \mathbf{M}]\mathbf{x}^{ML} + [\mathbf{I} - \mathbf{M}]\mathbf{x}^{ML}$$

$$= \mathbf{M}^2\mathbf{x}_0 + \mathbf{M}\mathbf{x}^{ML} - \mathbf{M}^2\mathbf{x}^{ML} + \mathbf{x}^{ML} - \mathbf{M}\mathbf{x}^{ML}$$

$$= \mathbf{M}^2\mathbf{x}_0 + \mathbf{x}^{ML} - \mathbf{M}^2\mathbf{x}^{ML} = \mathbf{M}^2\mathbf{x}_0 + [\mathbf{I} - \mathbf{M}^2]\mathbf{x}^{ML}$$

$$\boxed{\mathbf{x}_k = \mathbf{M}^k\mathbf{x}_0 + [\mathbf{I} - \mathbf{M}^k]\mathbf{x}^{ML}}$$

رابطه‌ی بین رگولاریزاسیون و توقف زودهنگام

تراجکتوری تندترین شیب

STEEPEST DESCENT TRAJECTORY

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \mathbf{g}_k = \mathbf{x}_k - \alpha(\mathbf{A}\mathbf{x}_k + \mathbf{d})$$

$$= \mathbf{x}_k - \alpha\mathbf{A}(\mathbf{x}_k + \mathbf{A}^{-1}\mathbf{d}) = \mathbf{x}_k - \alpha\mathbf{A}(\mathbf{x}_k - \mathbf{x}^{ML})$$

$$= [\mathbf{I} - \alpha\mathbf{A}]\mathbf{x}_k + \alpha\mathbf{A}\mathbf{x}^{ML} = \mathbf{M}\mathbf{x}_k + [\mathbf{I} - \mathbf{M}]\mathbf{x}^{ML}$$

$$\boxed{\mathbf{M} = [\mathbf{I} - \alpha\mathbf{A}]}$$

$$\mathbf{x}_1 = \mathbf{M}\mathbf{x}_0 + [\mathbf{I} - \mathbf{M}]\mathbf{x}^{ML}$$

$$\mathbf{x}_2 = \mathbf{M}\mathbf{x}_1 + [\mathbf{I} - \mathbf{M}]\mathbf{x}^{ML}$$

$$= \mathbf{M}^2\mathbf{x}_0 + \mathbf{M}[\mathbf{I} - \mathbf{M}]\mathbf{x}^{ML} + [\mathbf{I} - \mathbf{M}]\mathbf{x}^{ML}$$

$$= \mathbf{M}^2\mathbf{x}_0 + \mathbf{M}\mathbf{x}^{ML} - \mathbf{M}^2\mathbf{x}^{ML} + \mathbf{x}^{ML} - \mathbf{M}\mathbf{x}^{ML}$$

$$= \mathbf{M}^2\mathbf{x}_0 + \mathbf{x}^{ML} - \mathbf{M}^2\mathbf{x}^{ML} = \mathbf{M}^2\mathbf{x}_0 + [\mathbf{I} - \mathbf{M}^2]\mathbf{x}^{ML}$$

$$\boxed{\mathbf{x}_k = \mathbf{M}^k\mathbf{x}_0 + [\mathbf{I} - \mathbf{M}^k]\mathbf{x}^{ML}}$$

Regularization

$$F(\mathbf{x}) = E_D + \rho E_W \qquad (\rho = \alpha/\beta)$$

$$E_W = \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T(\mathbf{x} - \mathbf{x}_0)$$

To locate the minimum point, set the gradient to zero.

$$\nabla F(\mathbf{x}) = \nabla E_D + \rho \nabla E_W$$

$$\nabla E_W = (\mathbf{x} - \mathbf{x}_0) \qquad \nabla E_D = \mathbf{A}(\mathbf{x} - \mathbf{x}^{ML})$$

$$\nabla F(\mathbf{x}) = \mathbf{A}(\mathbf{x} - \mathbf{x}^{ML}) + \rho(\mathbf{x} - \mathbf{x}_0) = \mathbf{0}$$

رابطه‌ی بین رگولاریزاسیون و توقف زودهنگام

رگولاریزاسیون

REGULARIZATION

$$F(\mathbf{x}) = E_D + \rho E_W \qquad (\rho = \alpha/\beta)$$

$$E_W = \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T(\mathbf{x} - \mathbf{x}_0)$$

To locate the minimum point, set the gradient to zero.

$$\nabla F(\mathbf{x}) = \nabla E_D + \rho \nabla E_W$$

$$\nabla E_W = (\mathbf{x} - \mathbf{x}_0) \qquad \nabla E_D = \mathbf{A}(\mathbf{x} - \mathbf{x}^{ML})$$

$$\nabla F(\mathbf{x}) = \mathbf{A}(\mathbf{x} - \mathbf{x}^{ML}) + \rho(\mathbf{x} - \mathbf{x}_0) = \mathbf{0}$$

MAP – ML

$$\mathbf{A}(\mathbf{x}^{MP} - \mathbf{x}^{ML}) = -\rho(\mathbf{x}^{MP} - \mathbf{x}_0) = -\rho(\mathbf{x}^{MP} - \mathbf{x}^{ML} + \mathbf{x}^{ML} - \mathbf{x}_0)$$

$$= -\rho(\mathbf{x}^{MP} - \mathbf{x}^{ML}) - \rho(\mathbf{x}^{ML} - \mathbf{x}_0)$$

$$(\mathbf{A} + \rho\mathbf{I})(\mathbf{x}^{MP} - \mathbf{x}^{ML}) = \rho(\mathbf{x}_0 - \mathbf{x}^{ML})$$

$$(\mathbf{x}^{MP} - \mathbf{x}^{ML}) = \rho(\mathbf{A} + \rho\mathbf{I})^{-1}(\mathbf{x}_0 - \mathbf{x}^{ML})$$

$$\mathbf{x}^{MP} = \mathbf{x}^{ML} - \rho(\mathbf{A} + \rho\mathbf{I})^{-1}\mathbf{x}^{ML} + \rho(\mathbf{A} + \rho\mathbf{I})^{-1}\mathbf{x}_0 = \mathbf{x}^{ML} - \mathbf{M}_\rho\mathbf{x}^{ML} + \mathbf{M}_\rho\mathbf{x}_0$$

$$\mathbf{M}_\rho = \rho(\mathbf{A} + \rho\mathbf{I})^{-1}$$

$$\boxed{\mathbf{x}^{MP} = \mathbf{M}_\rho\mathbf{x}_0 + [\mathbf{I} - \mathbf{M}_\rho]\mathbf{x}^{ML}}$$

رابطه‌ی بین رگولاریزاسیون و توقف زودهنگام

ماکزیمم درست‌نمایی – ماکزیمم احتمال پسین

MAP – ML

$$\mathbf{A}(\mathbf{x}^{MP} - \mathbf{x}^{ML}) = -\rho(\mathbf{x}^{MP} - \mathbf{x}_0) = -\rho(\mathbf{x}^{MP} - \mathbf{x}^{ML} + \mathbf{x}^{ML} - \mathbf{x}_0)$$

$$= -\rho(\mathbf{x}^{MP} - \mathbf{x}^{ML}) - \rho(\mathbf{x}^{ML} - \mathbf{x}_0)$$

$$(\mathbf{A} + \rho\mathbf{I})(\mathbf{x}^{MP} - \mathbf{x}^{ML}) = \rho(\mathbf{x}_0 - \mathbf{x}^{ML})$$

$$(\mathbf{x}^{MP} - \mathbf{x}^{ML}) = \rho(\mathbf{A} + \rho\mathbf{I})^{-1}(\mathbf{x}_0 - \mathbf{x}^{ML})$$

$$\mathbf{x}^{MP} = \mathbf{x}^{ML} - \rho(\mathbf{A} + \rho\mathbf{I})^{-1}\mathbf{x}^{ML} + \rho(\mathbf{A} + \rho\mathbf{I})^{-1}\mathbf{x}_0 = \mathbf{x}^{ML} - \mathbf{M}_\rho\mathbf{x}^{ML} + \mathbf{M}_\rho\mathbf{x}_0$$

$$\mathbf{M}_\rho = \rho(\mathbf{A} + \rho\mathbf{I})^{-1}$$

$$\boxed{\mathbf{x}^{MP} = \mathbf{M}_\rho\mathbf{x}_0 + [\mathbf{I} - \mathbf{M}_\rho]\mathbf{x}^{ML}}$$

Early Stopping – Regularization

$$\boxed{\mathbf{x}_k = \mathbf{M}^k \mathbf{x}_0 + [\mathbf{I} - \mathbf{M}^k]\mathbf{x}^{ML}}$$

$$\boxed{\mathbf{x}^{MP} = \mathbf{M}_\rho \mathbf{x}_0 + [\mathbf{I} - \mathbf{M}_\rho]\mathbf{x}^{ML}}$$

$$\mathbf{M} = [\mathbf{I} - \alpha\mathbf{A}]$$

$$\mathbf{M}_\rho = \rho(\mathbf{A} + \rho\mathbf{I})^{-1}$$

Eigenvalues of $\mathbf{M}^k$:

$$[\mathbf{I} - \alpha\mathbf{A}]\mathbf{z}_i = \mathbf{z}_i - \alpha\mathbf{A}\mathbf{z}_i = \mathbf{z}_i - \alpha\lambda_i\mathbf{z}_i = \underbrace{(1 - \alpha\lambda_i)}_{\text{Eigenvalues of } \mathbf{M}}\mathbf{z}_i$$

$\mathbf{z}_i$ - eigenvector of $\mathbf{A}$

λ_i - eigenvalue of $\mathbf{A}$

$$\mathrm{eig}(\mathbf{M}^k) = (1 - \alpha\lambda_i)^k$$

Eigenvalues of $\mathbf{M}_\rho$:

$$\mathrm{eig}(\mathbf{M}_\rho) = \frac{\rho}{(\lambda_i + \rho)}$$

رابطه‌ی بین رگولاریزاسیون و توقف زودهنگام

توقف زودهنگام – رگولاریزاسیون

Early Stopping – Regularization

$$\boxed{\mathbf{x}_k = \mathbf{M}^k \mathbf{x}_0 + [\mathbf{I} - \mathbf{M}^k]\mathbf{x}^{ML}}$$

$$\mathbf{M} = [\mathbf{I} - \alpha \mathbf{A}]$$

$$\boxed{\mathbf{x}^{MP} = \mathbf{M}_\rho \mathbf{x}_0 + [\mathbf{I} - \mathbf{M}_\rho]\mathbf{x}^{ML}}$$

$$\mathbf{M}_\rho = \rho(\mathbf{A} + \rho\mathbf{I})^{-1}$$

Eigenvalues of $\mathbf{M}^k$:

$$[\mathbf{I} - \alpha\mathbf{A}]\mathbf{z}_i = \mathbf{z}_i - \alpha\mathbf{A}\mathbf{z}_i = \mathbf{z}_i - \alpha\lambda_i\mathbf{z}_i = \underbrace{(1 - \alpha\lambda_i)}_{\text{Eigenvalues of } \mathbf{M}}\mathbf{z}_i$$

$\mathbf{z}_i$ - eigenvector of $\mathbf{A}$

λ_i - eigenvalue of $\mathbf{A}$

$$\mathrm{eig}(\mathbf{M}^k) = (1 - \alpha\lambda_i)^k$$

Eigenvalues of $\mathbf{M}_\rho$:

$$\mathrm{eig}(\mathbf{M}_\rho) = \frac{\rho}{(\lambda_i + \rho)}$$

13 Reg. Parameter – Iteration Number

$\mathbf{M}^k$ and $\mathbf{M}_\rho$ have the same eigenvectors. They would be equal if their eigenvalues were equal.

$$\frac{\rho}{(\lambda_i + \rho)} = (1 - \alpha\lambda_i)^k$$

Taking log :

$$-\log\left(1 + \frac{\lambda_i}{\rho}\right) = k\log(1 - \alpha\lambda_i)$$

Since these are equal at $\lambda_i = 0$, they are always equal if the slopes are equal.

$$-\frac{1}{\left(1 + \frac{\lambda_i}{\rho}\right)}\frac{1}{\rho} = \frac{k}{1 - \alpha\lambda_i}(-\alpha) \quad \Longrightarrow \quad \alpha k = \frac{1}{\rho}\,\frac{(1 - \alpha\lambda_i)}{(1 + \lambda_i/\rho)}$$

If $\alpha\lambda_i$ and λ_i/ρ are small, then:

$$\boxed{\alpha k \cong \frac{1}{\rho}}$$

(Increasing the number of iterations is equivalent to decreasing the regularization parameter!)

رابطه‌ی بین رگولاریزاسیون و توقف زودهنگام

پارامتر رگولاریزاسیون – تعداد تکرار

REG. PARAMETER – ITERATION NUMBER

$\mathbf{M}^k$ and $\mathbf{M}_\rho$ have the same eigenvectors. They would be equal if their eigenvalues were equal.

$$\frac{\rho}{(\lambda_i + \rho)} = (1 - \alpha\lambda_i)^k \qquad \text{Taking log :} \qquad -\log\left(1 + \frac{\lambda_i}{\rho}\right) = k\log(1 - \alpha\lambda_i)$$

Since these are equal at $\lambda_i = 0$, they are always equal if the slopes are equal.

$$\text{مشتق} \Rightarrow \quad -\frac{1}{\left(1 + \frac{\lambda_i}{\rho}\right)}\frac{1}{\rho} = \frac{k}{1 - \alpha\lambda_i}(-\alpha) \quad \Longrightarrow \quad \alpha k = \frac{1}{\rho}\,\frac{(1 - \alpha\lambda_i)}{(1 + \lambda_i/\rho)}$$

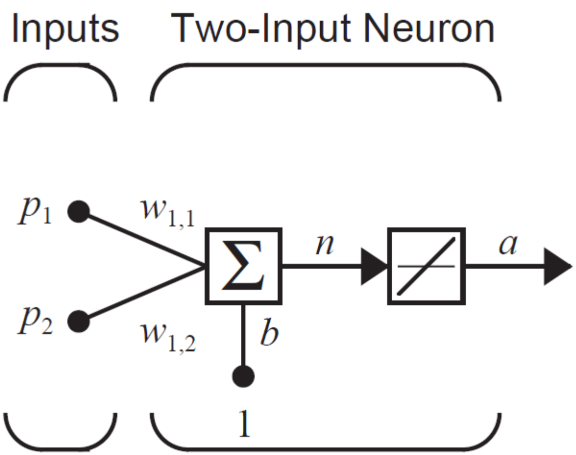
If $\alpha\lambda_i$ and λ_i/ρ are small, then:

$$\boxed{\alpha k \cong \frac{1}{\rho}}$$

افزایش تکرارها معادل با کاهش پارامتر رگولاریزاسیون است.

(Increasing the number of iterations is equivalent to decreasing the regularization parameter!)

Example



$$a = purelin(\mathbf{W}\mathbf{p} + b)$$

$$\left\{\mathbf{p}_1 = \begin{bmatrix}1\\1\end{bmatrix}, t_1 = 1\right\}$$ (Probability = 0.75)

$$\left\{\mathbf{p}_2 = \begin{bmatrix}-1\\1\end{bmatrix}, t_2 = -1\right\}$$ (Probability = 0.25)

$$F(\mathbf{x}) = E_D + \rho E_W$$

$$E_D = c + \mathbf{x}^T\mathbf{d} + \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}$$

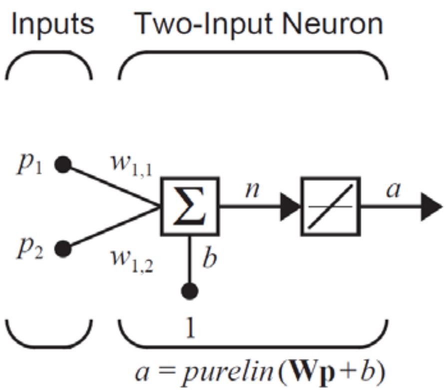
$$E_W = \frac{1}{2}\mathbf{x}^T\mathbf{x} \qquad c = 1 \qquad \mathbf{d} = -2\mathbf{h} = \begin{bmatrix}-2\\-1\end{bmatrix} \qquad \mathbf{A} = 2\mathbf{R} = \begin{bmatrix}2 & 1\\1 & 2\end{bmatrix}$$

$$\nabla^2 F(\mathbf{x}) = \nabla^2 E_D + \rho\nabla^2 E_W = \begin{bmatrix}2 & 1\\1 & 2\end{bmatrix} + \rho\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} = \begin{bmatrix}2+\rho & 1\\1 & 2+\rho\end{bmatrix}$$

رابطه‌ی بین رگولاریزاسیون و توقف زودهنگام

پارامتر رگولاریزاسیون – تعداد تکرار: مثال (۱ از ۴)

REG. PARAMETER – ITERATION NUMBER



$$\left\{\mathbf{p}_1 = \begin{bmatrix}1\\1\end{bmatrix}, t_1 = 1\right\}$$ (Probability = 0.75)

$$\left\{\mathbf{p}_2 = \begin{bmatrix}-1\\1\end{bmatrix}, t_2 = -1\right\}$$ (Probability = 0.25)

$$F(\mathbf{x}) = E_D + \rho E_W$$

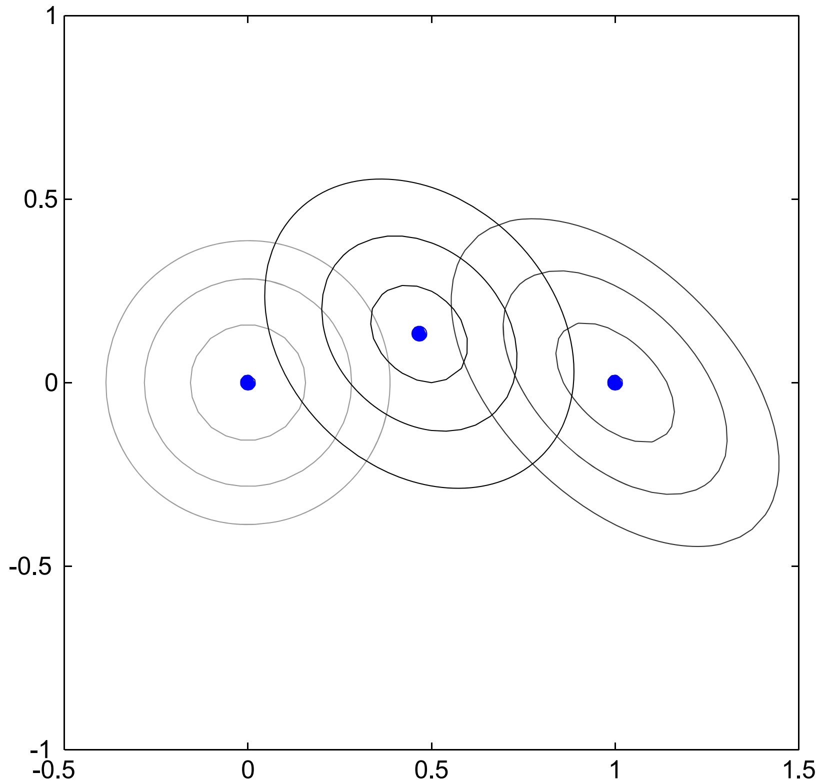
$$E_D = c + \mathbf{x}^T\mathbf{d} + \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}$$

$$E_W = \frac{1}{2}\mathbf{x}^T\mathbf{x} \qquad c = 1 \qquad \mathbf{d} = -2\mathbf{h} = \begin{bmatrix}-2\\-1\end{bmatrix} \qquad \mathbf{A} = 2\mathbf{R} = \begin{bmatrix}2 & 1\\1 & 2\end{bmatrix}$$

$$\nabla^2 F(\mathbf{x}) = \nabla^2 E_D + \rho\nabla^2 E_W = \begin{bmatrix}2 & 1\\1 & 2\end{bmatrix} + \rho\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} = \begin{bmatrix}2+\rho & 1\\1 & 2+\rho\end{bmatrix}$$

13

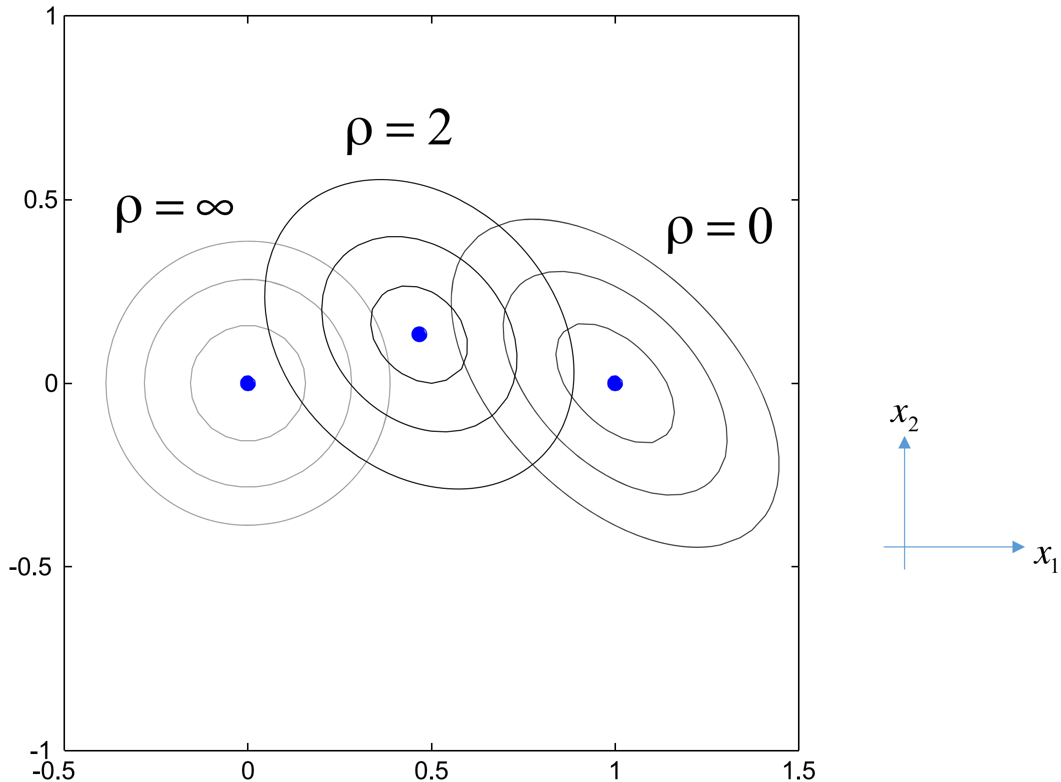
$\rho = 0, 2, \infty$



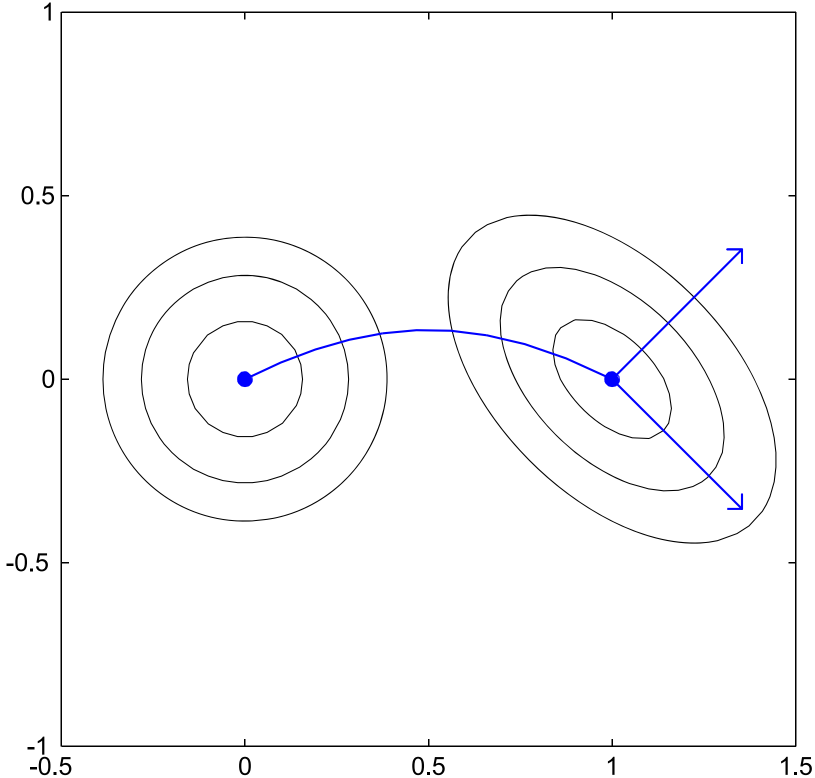
رابطه‌ی بین رگولاریزاسیون و توقف زودهنگام

پارامتر رگولاریزاسیون – تعداد تکرار: مثال (۲ از ۴)

Reg. Parameter – Iteration Number



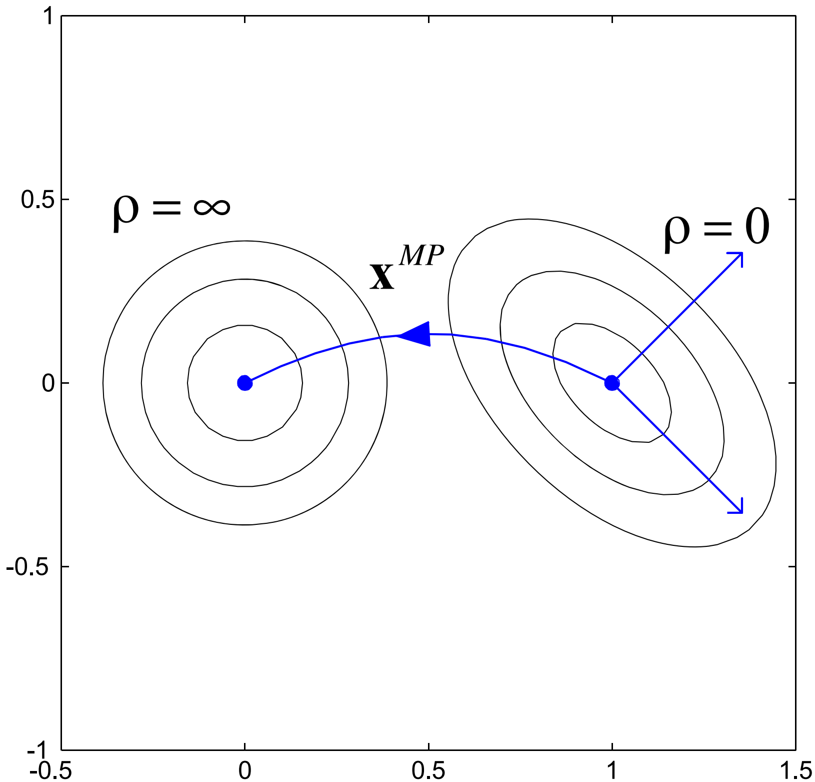
$\rho = 0 \to \infty$



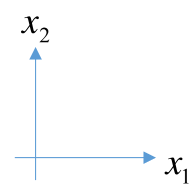
رابطه‌ی بین رگولاریزاسیون و توقف زودهنگام

پارامتر رگولاریزاسیون – تعداد تکرار: مثال (۳ از ۴)

Reg. Parameter – Iteration Number

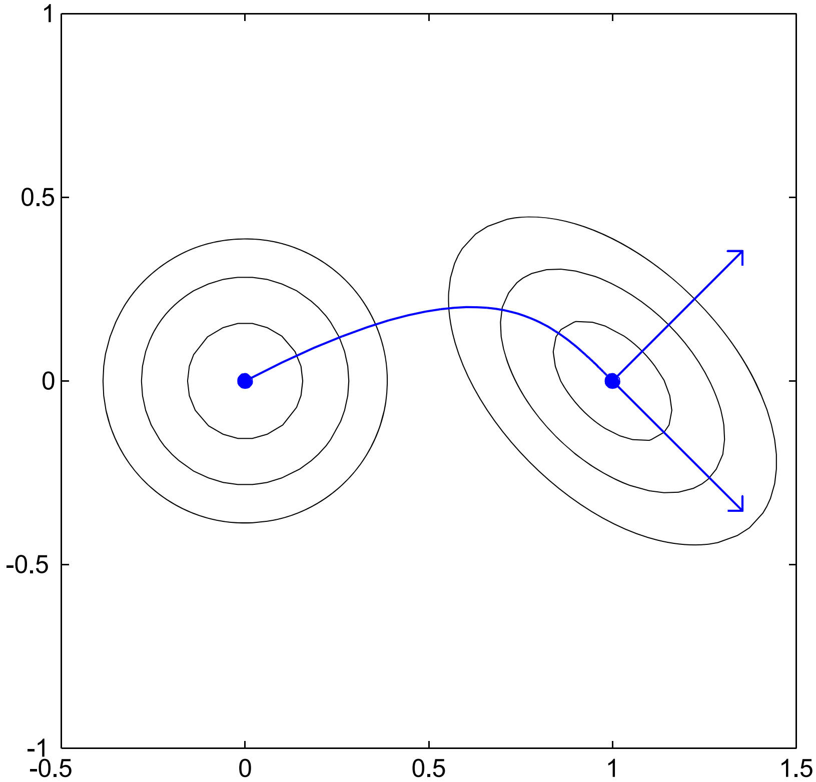


$\mathbf{x}^{MP}$ با تغییر ρ



افزایش تکرارها
معادل با کاهش ρ است.

Steepest Descent Path



رابطه‌ی بین رگولاریزاسیون و توقف زودهنگام

پارامتر رگولاریزاسیون – تعداد تکرار: مثال (۴ از ۴)

REG. PARAMETER – ITERATION NUMBER

نتیجه‌ی SD با توقف زودهنگام، مشابه (بسیار نزدیک به) نتیجه‌ی رگولاریزاسیون است.

تعداد تکرار بسیار کم

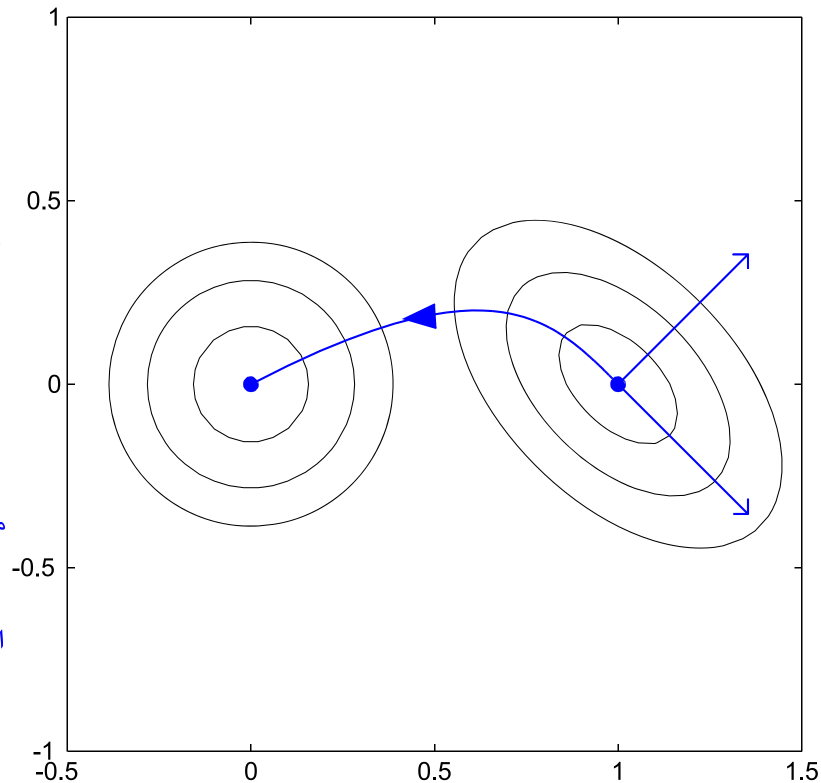
⇓

معادل با

مقدار بسیار بزرگ برای ρ است.

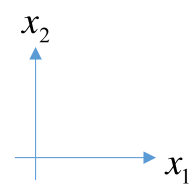
⇓

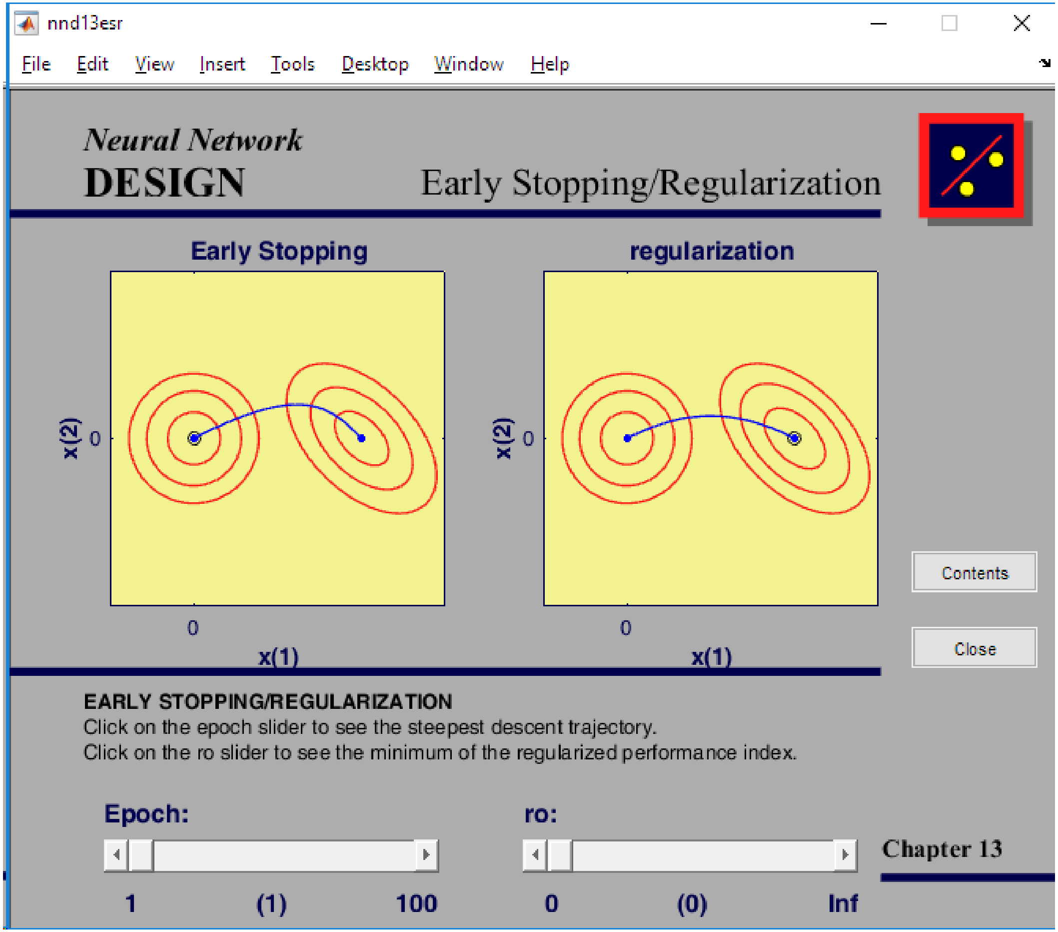
افزایش تعداد تکرار معادل با کاهش ρ است.



$$\rho = \frac{\alpha}{\beta}$$

پارامتر رگولاریزاسیون





MATLAB

```
>> nnd13esr
```

Effective Number of Parameters

$$\gamma = n - 2\alpha^{MP} \operatorname{tr}\left\{ (\mathbf{H}^{MP})^{-1} \right\}$$

$$\mathbf{H}(\mathbf{x}) = \nabla^2 F(\mathbf{x}) = \beta \nabla^2 E_D + \alpha \nabla^2 E_W = \beta \nabla^2 E_D + 2\alpha \mathbf{I}$$

$$\operatorname{tr}\{\mathbf{H}^{-1}\} = \sum_{i=1}^{n} \frac{1}{\beta\lambda_i + 2\alpha}$$

$$\gamma = n - 2\alpha^{MP} \operatorname{tr}\left\{ (\mathbf{H}^{MP})^{-1} \right\} = n - \sum_{i=1}^{n} \frac{2\alpha}{\beta\lambda_i + 2\alpha} = \sum_{i=1}^{n} \frac{\beta\lambda_i}{\beta\lambda_i + 2\alpha}$$

Effective number of parameters will equal number of large eigenvalues of the Hessian.

$$\gamma = \sum_{i=1}^{n} \frac{\beta\lambda_i}{\beta\lambda_i + 2\alpha} = \sum_{i=1}^{n} \gamma_i \qquad \gamma_i = \frac{\beta\lambda_i}{\beta\lambda_i + 2\alpha} \qquad 0 \le \gamma_i \le 1$$

تعداد مؤثر پارامترها

Effective Number of Parameters

$$\gamma = n - 2\alpha^{MP}\mathrm{tr}\left\{(\mathbf{H}^{MP})^{-1}\right\}$$

$$\mathbf{H}(\mathbf{x}) = \nabla^2 F(\mathbf{x}) = \beta\nabla^2 E_D + \alpha\nabla^2 E_W = \beta\nabla^2 E_D + 2\alpha\mathbf{I}$$

$$\mathrm{tr}\{\mathbf{H}^{-1}\} = \sum_{i=1}^{n}\frac{1}{\beta\lambda_i + 2\alpha}$$

$$\gamma = n - 2\alpha^{MP}\mathrm{tr}\left\{(\mathbf{H}^{MP})^{-1}\right\} = n - \sum_{i=1}^{n}\frac{2\alpha}{\beta\lambda_i + 2\alpha} = \sum_{i=1}^{n}\frac{\beta\lambda_i}{\beta\lambda_i + 2\alpha}$$

Effective number of parameters will equal number of large eigenvalues of the Hessian.

$$\gamma = \sum_{i=1}^{n}\frac{\beta\lambda_i}{\beta\lambda_i + 2\alpha} = \sum_{i=1}^{n}\gamma_i \qquad \gamma_i = \frac{\beta\lambda_i}{\beta\lambda_i + 2\alpha} \qquad 0 \le \gamma_i \le 1$$

مقدار λ_i بزرگ‌تر $\Leftarrow$ انحنای بیشتر در راستای $\mathbf{v}_i$

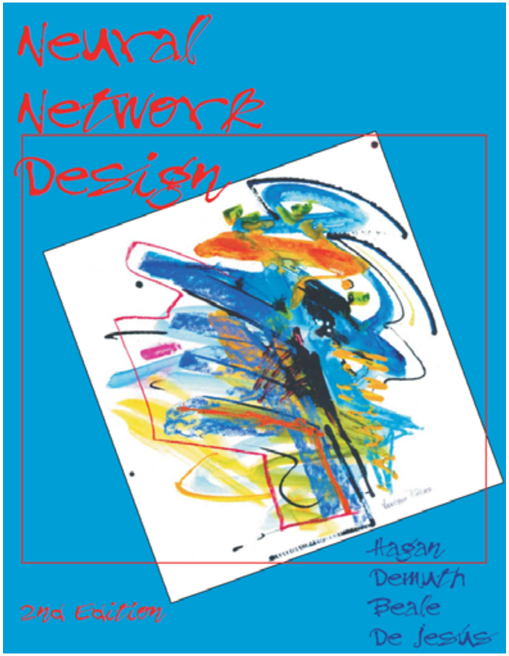
تعداد مؤثر پارامترها = تعداد مقادیر ویژه‌ی بزرگ (دور از صفر) ماتریس هسی $\nabla^2 E_D(\mathbf{x})$ است.

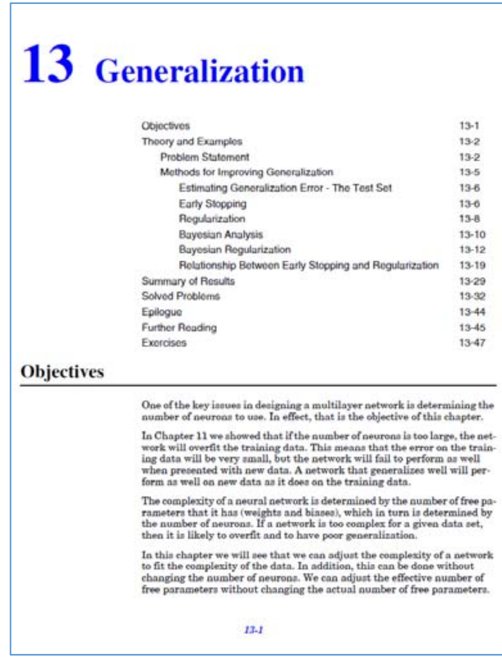
تعمیم

۹

منابع

منبع اصلی





13 Generalization

Objectives

One of the key issues in designing a multilayer network is determining the number of neurons to use. In effect, that is the objective of this chapter.

In Chapter 11 we showed that if the number of neurons is too large, the network will overfit the training data. This means that the error on the training data will be very small, but the network will fail to perform as well when presented with new data. A network that generalizes well will perform as well on new data as it does on the training data.

The complexity of a neural network is determined by the number of free parameters that it has (weights and biases), which in turn is determined by the number of neurons. If a network is too complex for a given data set, then it is likely to overfit and to have poor generalization.

In this chapter we will see that we can adjust the complexity of a network to fit the complexity of the data. In addition, this can be done without changing the number of neurons. We can adjust the effective number of free parameters without changing the actual number of free parameters.

13-1

Martin T. Hagan, Howard B. Demuth, Mark H. Beale, Orlando De Jesus,
Neural Network Design,
2nd Edition, Martin Hagan, 2014.
Chapter 13

Online version can be downloaded from: http://hagan.okstate.edu/nnd.html