

Υπερπροσαρμογή και Ομαλοποίηση (Overfitting & Regularization)

Μηχανική Μάθηση

ΔΠΜΣ Επιστήμης Δεδομένων & Μηχανικής Μάθησης

Γιώργος Αλεξανδρίδης – gealexandri@islab.ntua.gr

Εισαγωγικές Έννοιες

Προσαρμογή

Γενίκευση

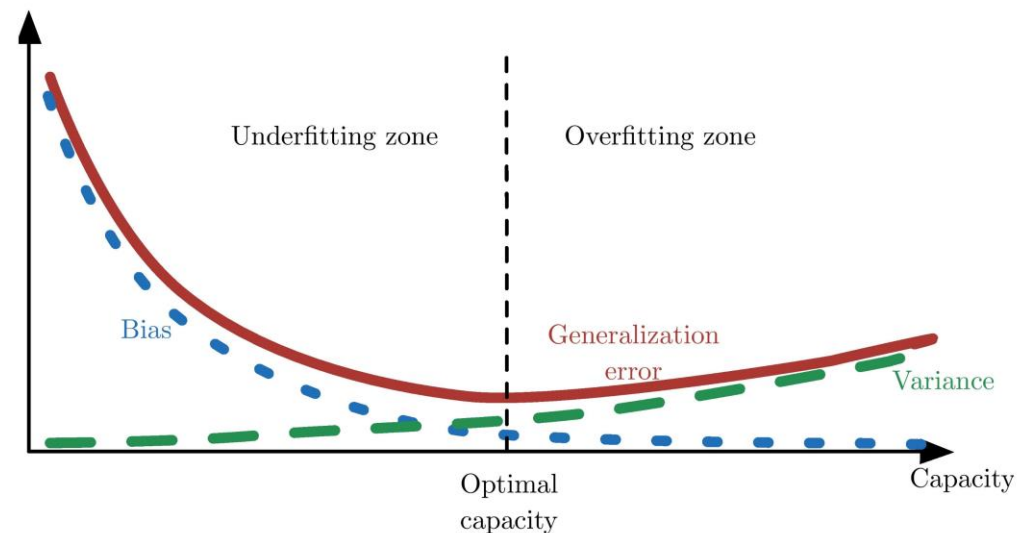
- **Επιθυμητό αποτέλεσμα** κάθε διαδικασίας μάθησης είναι το μοντέλο να έχει **καλή απόδοση** **όχι μόνο** στα δεδομένα που έχει **εκπαιδευτεί** αλλά και σε **νέα δεδομένα**, μια διαδικασία που είναι γνωστή ως **γενίκευση** (*generalization*)
 - Μικρό **σφάλμα εκπαίδευσης** (*training error*)
 - Μικρή διαφορά μεταξύ σφάλματος εκπαίδευσης και **σφάλματος γενίκευσης** (ή αλλιώς **σφάλματος ελέγχου** – *test error*)

Στατιστική Θεωρία μάθησης

- *Statistical Learning Theory*
- Βασική υπόθεση εργασίας
- Κάθε πρόβλημα μοντελοποιείται ως μια **διαδικασία παραγωγής δεδομένων** που προκύπτουν από **άγνωστη** σε εμάς **κατανομή**
- Κάθε πιθανό δείγμα των δεδομένων, είτε ανήκει στο σύνολο εκπαίδευσης είτε στο σύνολο ελέγχου, παράγεται **ανεξάρτητα** (*independent*) από τα υπόλοιπα, από την **ίδια κατανομή** (*identically distributed*) \Rightarrow *i.i.d assumptions*

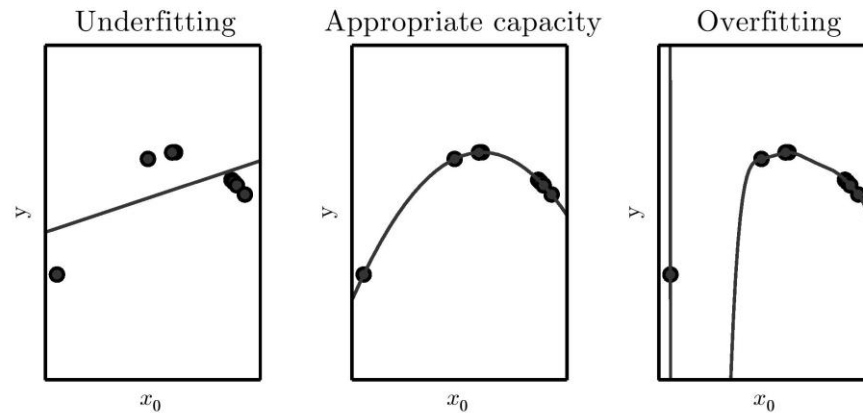
Υπερπροσαρμογή και Υποπροσαρμογή

- **Υποπροσαρμογή (*underfitting*)**
 - Το μοντέλο δεν μπορεί να μειώσει το **σφάλμα εκπαίδευσης**
- **Υπερπροσαρμογή (*overfitting*)**
 - Το μοντέλο δεν μπορεί να μειώσει το «**χάσμα**» μεταξύ **σφάλματος εκπαίδευσης και ελέγχου**
- **Χωρητικότητα (*capacity*)**
 - Καθορίζει την ικανότητα του μοντέλου να γενικεύει
 - Σχετίζεται άμεσα με τον **χώρο υποθέσεων (*hypothesis space*)** του μοντέλου



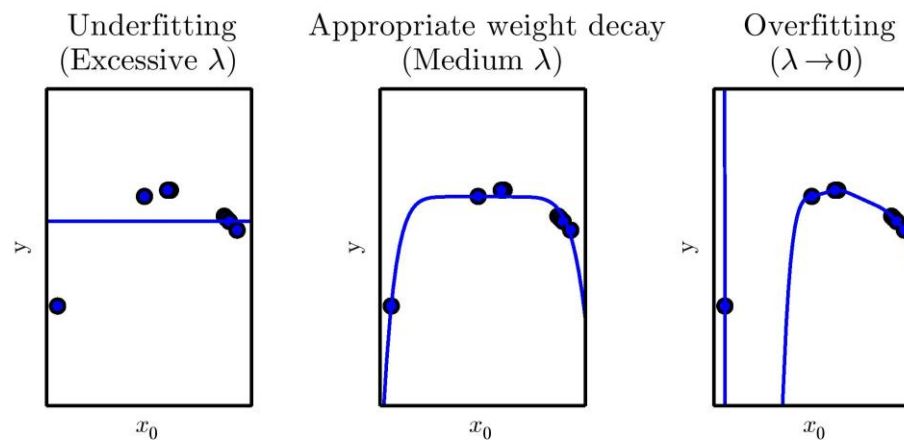
Χωρητικότητα και Δυνατότητα Μάθησης

- Τα μοντέλα μηχανικής μάθησης έχουν τη **βέλτιστη απόδοση** όταν η **χωρητικότητά** τους είναι η **κατάλληλη** για το **πρόβλημα** που τους ζητείται να μάθουν
- Τα μοντέλα με **ανεπαρκή χωρητικότητα** **δεν μπορούν** να επιλύσουν πολύπλοκα προβλήματα
- Τα μοντέλα με **υψηλή χωρητικότητα** **επιλύουν** πολύπλοκα προβλήματα, αλλά ενδέχεται να εμφανίσουν **υπερπροσαρμογή**



Έλεγχος του χώρου υποθέσεων

- Στις περισσότερες περιπτώσεις είναι **δύσκολο** να καθοριστεί επακριβώς εκ των προτέρων ο **χώρος υποθέσεων** ενός μοντέλου
- Μπορούμε, ωστόσο να ξεκινήσουμε από ένα μεγάλο χώρο υποθέσεων και όσο προχωράει η εκπαίδευση να **ενισχύουμε** ή να **εξασθενούμε** συγκεκριμένα χαρακτηριστικά

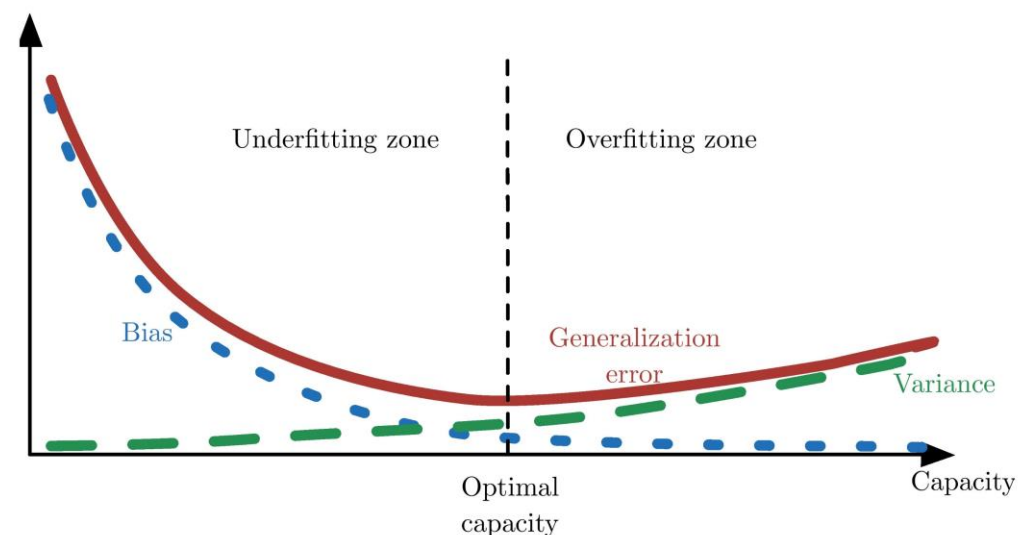


Μεροληψία και Διακύμανση

- Μετρούν δύο διαφορετικά είδη σφάλματος σε έναν εκτιμητή
- **Μεροληψία** (*bias*)
 - Μέτρο της αναμενόμενης απόκλισης της πρόβλεψης του εκτιμητή από την πραγματική τιμή
 - **Υψηλή μεροληψία** σημαίνει ότι το μοντέλο έχει «**υπεραπλουστεύσει**» τα δεδομένα
- **Διακύμανση** (*variance*)
 - Μέτρο της απόκλισης των προβλέψεων του εκτιμητή σε «παραπλήσια» δεδομένα
 - **Υψηλή διακύμανση** σημαίνει ότι το μοντέλο δεν μπορεί να **γενικεύσει**

Ισορροπία μεταξύ μεροληψίας και διακύμανσης

- Bias-variance trade-off
- Στόχος κάθε διαδικασίας μάθησης
- Στόχος της **ομαλοποίησης** είναι να **μειώσουμε** τη διακύμανση του μοντέλου, **χωρίς** να **επιηρεαστεί** η **μεροληψία** του



Εισαγωγικές Έννοιες

Ομαλοποίηση

Ομαλοποίηση

- Ορισμός
 - *Κάθε τροποποίηση ενός αλγορίθμου μάθησης που στόχο έχει τη μείωση του σφάλματος γενίκευσης του και όχι του σφάλματος εκπαίδευσής του καλείται ομαλοποίηση*
- Τεχνικές
 - Προσθήκη περιορισμών στο μοντέλο
 - Επιπλέον όροι στην αντικειμενική συνάρτηση
 - Εισαγωγή εκ των προτέρων γνώσης
 - Προτίμηση απλούστερων μοντέλων
 - Μέθοδοι ensemble
 - Συνδυασμός πολλαπλών υποθέσεων για την περιγραφή των δεδομένων εκπαίδευσης

Όροι πεινής

Όροι ποινής (penalties)

- Μείωση χωρητικότητας μοντέλου μέσω προσθήκης **όρου ποινής** Ω στην αντικειμενική συνάρτηση J
- Από τις πρώτες προσεγγίσεις ομαλοποίησης στο χώρο της μηχανικής μάθησης
- Πρόβλημα ταξινόμησης: $\tilde{J}(\theta; \mathbf{X}, \mathbf{y}) = J(\theta; \mathbf{X}, \mathbf{y}) + \alpha\Omega(\theta)$
 - \tilde{J} ομαλοποιημένη αντικειμενική συνάρτηση
 - θ χώρος παραμέτρων, \mathbf{X} δεδομένα εκπαίδευσης, \mathbf{y} ετικέτες εξόδου
 - α **υπερπαραμέτρος** που ρυθμίζει τη **συνεισφορά** του παράγοντα της **ομαλοποίησης** στην αντικειμενική συνάρτηση
 - Για $\alpha = 0$ δεν έχουμε **καθόλου** ομαλοποίηση
 - Όσο το α μεγαλώνει, η ομαλοποίηση παίζει μεγαλύτερο ρόλο
- Στην πράξη, η ομαλοποίηση αφορά τα **βάρη** και όχι τις **πολώσεις**
 - Ομαλοποίηση στις πολώσεις μπορεί να οδηγήσει σε προβλήματα *υποπροσαρμογής*
 - $\tilde{J}(w; \mathbf{X}, \mathbf{y}) = J(w; \mathbf{X}, \mathbf{y}) + \alpha\Omega(w)$

Όροι ποινηής

Ομαλοποίηση L^2

Ομαλοποίηση L^2

- Εναλλακτικές ονομασίες
 - Φθορά βαρών (*weight decay*), **αμφικλινής παλινδρόμηση** (*ridge regression*), **κανονικοποίηση Tikhonov** (*Tikhonov regularization*)
- Προσθήκη όρου ομαλοποίησης βαρών $\Omega(\boldsymbol{\theta}) = \frac{1}{2} \|\mathbf{w}\|_2^2 = \frac{1}{2} \mathbf{w}^T \mathbf{w}$ στην J
 - $\tilde{J}(\mathbf{w}; \mathbf{X}, \mathbf{y}) = J(\mathbf{w}; \mathbf{X}, \mathbf{y}) + \alpha \frac{1}{2} \mathbf{w}^T \mathbf{w}$
- Ενημέρωση βαρών μέσω στοχαστικής κατάβασης κλίσης (SGD)
 - Υπολογισμός κλίσης
 - $\nabla_{\mathbf{w}} \tilde{J}(\mathbf{w}; \mathbf{X}, \mathbf{y}) = \nabla_{\mathbf{w}} J(\mathbf{w}; \mathbf{X}, \mathbf{y}) + \alpha \mathbf{w}$
 - Ενημέρωση βαρών
 - $\mathbf{w}^{\tau+1} \leftarrow \mathbf{w}^{\tau} - \epsilon (\nabla_{\mathbf{w}^{\tau}} J(\mathbf{w}^{\tau}; \mathbf{X}, \mathbf{y}) + \alpha \mathbf{w}^{\tau}) \Rightarrow \mathbf{w}^{\tau+1} \leftarrow (1 - \epsilon \alpha) \mathbf{w}^{\tau} - \epsilon \nabla_{\mathbf{w}^{\tau}} J(\mathbf{w}^{\tau}; \mathbf{X}, \mathbf{y})$
 - Το αποτέλεσμα είναι να **μειώνεται** («φθεΐρεται») το εύρος του διανύσματος των βαρών κατά παράγοντα σε κάθε βήμα.

Ομαλοποίηση L^2 : Επίδραση στη διαδικασία μάθησης (1/3)

- Έστω w^* το διάνυσμα βαρών που ελαχιστοποιεί την J
- Αντικατάσταση της J από την τετραγωνική της προσέγγιση \hat{J} γύρω από το w^*
 - Ειδικά για προβλήματα γραμμικής παλινδρόμησης (*linear regression*) όπου η αντικειμενική συνάρτηση υπολογίζει διαφορές τετραγώνων (π.χ. MSE), η προσέγγιση είναι τέλεια
 - $\hat{J}(w) = J(w^*) + \frac{1}{2} (w - w^*)^T H (w - w^*)$
 - Δεν υπάρχει πρωτοβάθμιος όρος μιας εξ' ορισμού είναι ελάχιστο και άρα η κλίση γύρω από το $w - w^*$ είναι (σχεδόν) μηδενική
 - H : Εσσιανός Πίνακας (*Hessian Matrix*) όλων των $\frac{\partial^2 J}{\partial w_i \partial w_j}$
 - Επειδή βρισκόμαστε γύρω από ελάχιστο, ο H είναι θετικά ημι-καθορισμένος

Ομαλοποίηση L^2 : Επίδραση στη διαδικασία μάθησης (2/3)

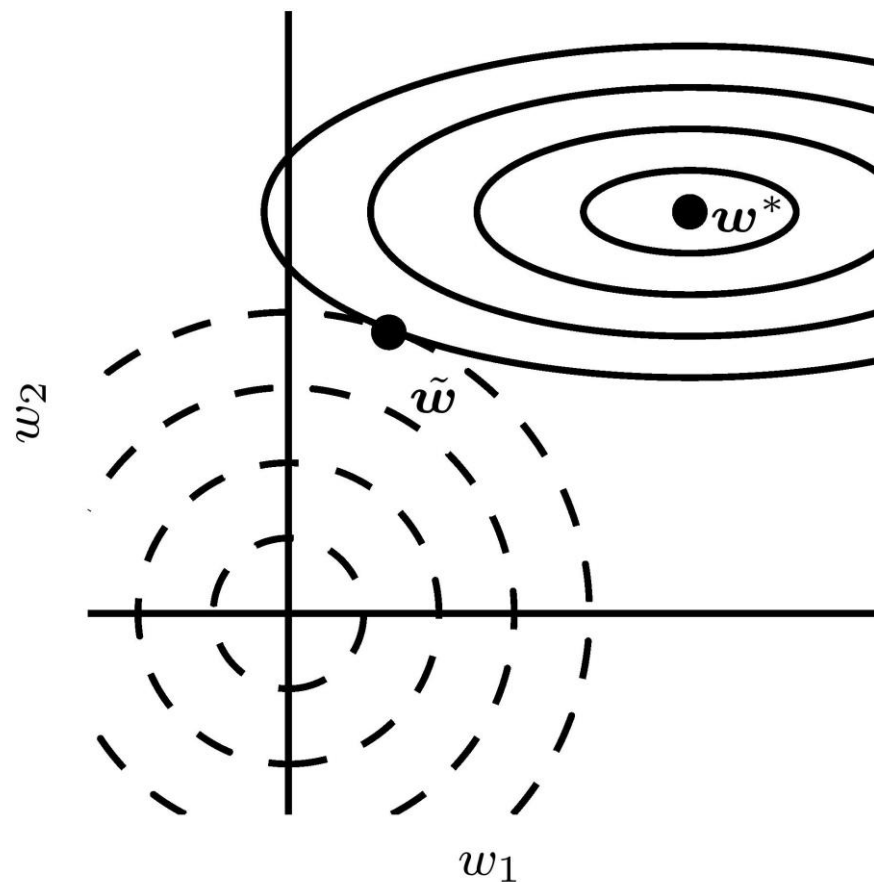
- Η \hat{J} ελαχιστοποιείται στα σημεία όπου η κλίση της $\nabla_w \hat{J}(w; \mathbf{X}, \mathbf{y}) = \mathbf{H}(w - w^*)$ γίνεται ίση με το 0
- Προσθέτοντας τον όρο ποινής $\alpha \frac{1}{2} w^T w$ το τοπικό ελάχιστο αλλάζει και γίνεται πλέον \tilde{w}
 - $\nabla_w \left(\hat{J}(\tilde{w}) + \frac{\alpha}{2} \tilde{w}^T \tilde{w} \right) = 0 \Rightarrow \mathbf{H}(\tilde{w} - w^*) + \alpha \tilde{w} = 0 \Rightarrow \tilde{w} = (\mathbf{H} + \alpha \mathbf{I})^{-1} \mathbf{H} w^*$
- Όταν $\alpha \rightarrow 0$, το \tilde{w} προσεγγίζει το w^*
- Όταν $\alpha \neq 0$, χρησιμοποιούμε αποσύνθεση ιδιοτιμών (eigen decomposition) $\mathbf{H} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T$
 - Εφικτό μιας και \mathbf{H} συμμετρικός με πραγματικές τιμές
 - $\mathbf{\Lambda}$ διαγώνιος πίνακας ιδιοτιμών, \mathbf{Q} ορθο-κανονικός πίνακας ιδιοδιανυσμάτων
 - $\tilde{w} = \mathbf{Q}(\mathbf{\Lambda} + \alpha \mathbf{I})^{-1} \mathbf{\Lambda} \mathbf{Q}^T w^*$

Ομαλοποίηση L^2 : Επίδραση στη διαδικασία μάθησης (3/3)

- Ουσιαστικά το w^* **αναπροσαρμόζεται** στην κατεύθυνση των αξόνων που ορίζονται από τα ιδιοδιανύσματα του H κατά ένα **παράγοντα** $\frac{\lambda_i}{\lambda_i + \alpha}$
 - Για μεγάλες ιδιοτιμές ($\lambda_i \gg \alpha$), η επίδραση της ομαλοποίησης είναι **πολύ μικρή**
 - Για μικρές ιδιοτιμές ($\lambda_i \ll \alpha$), η επίδραση της ομαλοποίησης είναι **πολύ μεγάλη**
 - Συρρικνώνει την επίδραση των αντίστοιχων ιδιοδιανυσμάτων στο 0

Ομαλοποίηση L^2 : 1^ο Παράδειγμα

- **Συνεχείς** καμπύλες: **Ίδιες** τιμές J
- **Διακεκομμένες** καμπύλες: **Ίδιες** τιμές \hat{J}
- **Σημείο ισορροπίας**: \tilde{w}
- 1^η διάσταση: Ιδιοτιμή του \mathbf{H} χαμηλή
 - Η τιμή της J δεν μεταβάλλεται σημαντικά σε αυτό τον άξονα.
 - Επίδραση ομαλοποιητής **σημαντική**
 - «Τραβάει» το w_1 προς το 0
- 2^η διάσταση: Ιδιοτιμή του \mathbf{H} υψηλή
 - Η τιμή της J είναι **«ευαίσθητη»** σε **μικρές μεταβολές** του w_2
 - Επίδραση ομαλοποιητή **μικρή**



Ομαλοποίηση L^2 : 2^ο Παράδειγμα

- Γραμμική Παλινδρόμηση
 - **Συνάρτηση κόστους**: άθροισμα τετραγώνων σφάλματος:
 - $J(\mathbf{w}) = (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y})$
 - Βέλτιστο διάνυσμα βαρών: $\mathbf{w}^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$
 - Προσθήκη όρου ομαλοποίησης L^2
 - $J(\mathbf{w}) = (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y}) + \frac{\alpha}{2} \mathbf{w}^T \mathbf{w}$
 - Βέλτιστο διάνυσμα βαρών: $\tilde{\mathbf{w}} = (\mathbf{X}^T \mathbf{X} + \alpha \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$
- Ο όρος $\mathbf{X}^T \mathbf{X}$ είναι **ανάλογος** του **πίνακα συνδιασποράς** των χαρακτηριστικών της εισόδου
 - **Διαγώνιες τιμές**: Αντιστοιχούν στη διακύμανση των χαρακτηριστικών της εισόδου
 - Η ομαλοποίηση L^2 **αναγκάζει** τον αλγόριθμο μάθησης να θεωρήσει ότι η είσοδος παρουσιάζει **μεγαλύτερη διακύμανση** και συνεπώς να **μικρύνει** τα **βάρη** εκείνα που εμφανίζουν **μικρότερη συνδιασπορά**.

Όροι ποινηής

Ομαλοποίηση L^1

Ομαλοποίηση L^1

- Προσθήκη όρου ομαλοποίησης βαρών $\Omega(\boldsymbol{\theta}) = \|\mathbf{w}\|_1$ στην J
 - Ίσο με το άθροισμα των απόλυτων τιμών των βαρών
- Ομαλοποιημένη αντικειμενική συνάρτηση
 - $\tilde{J}(\mathbf{w}; \mathbf{X}, \mathbf{y}) = J(\mathbf{w}; \mathbf{X}, \mathbf{y}) + \alpha\|\mathbf{w}\|_1$
- Υπολογισμός κλίσης: $\nabla_{\mathbf{w}} \tilde{J}(\mathbf{w}; \mathbf{X}, \mathbf{y}) = \nabla_{\mathbf{w}} J(\mathbf{w}; \mathbf{X}, \mathbf{y}) + \alpha \text{sgn}(\mathbf{w})$
- **Διαφορά** ως προς ομαλοποίηση L^2
 - Η συνεισφορά της ομαλοποίησης στην κλίση **δεν** είναι πλέον **ανάλογη** του εύρους w_i της κάθε παραμέτρου, αλλά του προσήμου της
 - Για να προχωρήσουμε, κάνουμε την **επιπλέον παραδοχή** ότι ο εσσιανός πίνακας είναι **διαγώνιος** με $H_{i,i} > 0$
 - Έχουν αφαιρεθεί οι συσχετίσεις μεταξύ των χαρακτηριστικών της εισόδου λ.χ. με την προεπεξεργασία μέσω PCA

Ομαλοποίηση L^1 : Χαρακτηριστικά

- Τετραγωνική προσέγγιση \hat{J} της J
 - $\hat{J}(\mathbf{w}) = J(\mathbf{w}^*) + \sum_i \left[\frac{1}{2} H_{i,i} (\mathbf{w} - \mathbf{w}^*)^2 + \alpha |w_i| \right]$
- Τοπικό ελάχιστο: $\tilde{w}_i = \text{sgn}(w_i^*) \max \left(|w_i^*| - \frac{\alpha}{H_{i,i}}, 0 \right)$
 - Όταν $|w_i^*| \leq \frac{\alpha}{H_{i,i}}$, τότε το \tilde{w}_i γίνεται 0
 - Όταν $|w_i^*| > \frac{\alpha}{H_{i,i}}$, τότε το \tilde{w}_i «σύρεται» προς το 0 κατά έναν όρο $\frac{\alpha}{H_{i,i}}$
- Η ομαλοποίηση L^1 οδηγεί σε πιο **αραιές** (*sparse*) αναπαραστάσεις σε σύγκριση με την L^2
 - Υπό την έννοια ότι **περισσότερες παράμετροι** έχουν **μηδενικές τιμές**
 - Χρησιμοποιείται ως μηχανισμός **επιλογής χαρακτηριστικών** (*feature selection*)

Χαρακτηριστικά L^1 και L^2 ομαλοποίησης

- Είναι ισοδύναμες με την **εκτίμηση της μέγιστης εκ των υστέρων πιθανότητας** (MAP) της *Μπεϋζιανής συμπερασματολογίας*
- Ομαλοποίηση L^1
 - Εκ των προτέρων πιθανότητα ακολουθεί μια **ισοτροπική κατανομή Laplace**
- Ομαλοποίηση L^2
 - Εκ των προτέρων πιθανότητα ακολουθεί μια **κανονική κατανομή**

Βιβλιογραφία

- Πηγή: *Ian Goodfellow, Yoshua Bengio, Aaron Courville* “Deep Learning” – MIT Press (<https://www.deeplearningbook.org/>)
- Εισαγωγικές Έννοιες
 - Χωρητικότητα, Υπερπροσαρμογή, υποπροσαρμογή (§5.2, §5.2.1)
 - Ισορροπία Μεροληψίας και Διακύμανσης (§5.4.4)
 - Προσαρμογή (§5.2.2, §7)
- Ομαλοποίηση μέσω της προσθήκης όρων ποινής (§7.1)
 - Ομαλοποίηση L^2 (§7.1.1)
 - Ομαλοποίηση L^1 (§7.1.2)