Αναγνώριση Προτύπων - Μηχανική Μάθηση

Μέθοδοι κανονικοποίησης (regularization methods)

Γιάννης Παναγάκης

Επισκόπηση

- Σε αυτό το μάθημα θα συζητήσουμε αρχικά την έννοια της κανονικοποίησης (regularization) στους αλγορίθμους μηχανικής μάθησης.

- Θα δούμε πώς σχεδιάζουμε διαφορετικούς αλγορίθμους μηχανικής μάθησης επιλέγοντας κατάλληλες συναρτήσεις κανονικοποιησης (regularizers) και κόστους (loss functions).

Regularization

- Γιατί υπάρχει χάσμα γενίκευσης μεταξύ των δεδομένων εκπαίδευσης και ελέγχου;
  - Overfittting (το μοντέλο περιγράφει στατιστικές ιδιαιτερότητες)
  - Το μοντέλο δεν περιορίζεται σε περιοχές όπου δεν υπάρχουν παραδείγματα εκπαίδευσης
- Κανονικοποίηση = μέθοδοι για τη μείωση του χάσματος γενίκευσης
- Τεχνικά σημαίνει την προσθήκη όρων στη συνάρτηση απώλειας

Ελαχιστοποίηση εμπειρικού σφάλματος

- Η ελαχιστοποίηση του εμπειρικού σφάλματος (empirical risk minimisation - ERM) αποτελεί την πιο διαδεδομένη προσέγγιση για τη σχεδίαση αλγορίθμων μηχανικής μάθησης.

- Θυμηθείτε ότι το πραγματικό ρίσκο δεν υπολογίζεται καθώς η p(x,y) είναι άγνωστη.

$$R(f) = E_{(\mathbf{x},y)\sim p}\left[\ell\left(y, f(\mathbf{x})\right)\right] = \iint_{x\,y} \ell\left(y, f(\mathbf{x})\right) p(\mathbf{x}, y)\partial y \partial \mathbf{x}\,.$$

- Η συνάρτηση κόστους (loss function) ποσοτικοποιεί το κόστος της πρόβλεψης *f(x)* στη θέση του (πραγματικού) *y* και συμβολίζεται με $\ell(y, f(x))$

- Γενική ιδέα: Θωρούμε ότι το εμπειρικό ρίσκο $\hat{R}(f)$ υποκαθιστά το πραγματικό ρίσκο

$$\hat{R}(f) = E_{(\mathbf{x},y)\sim D}\left[\ell\left(y, f(\mathbf{x})\right)\right] = \frac{1}{N}\sum_{i=1}^{N} \ell\left(y_i, f(\mathbf{x}_i)\right)$$

- Σύνολο εκπαίδευσης: $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N} \sim p$

Ελαχιστοποίηση εμπειρικού σφάλματος

- Για να μεταβούμε από την παραπάνω γενική ιδέα σε αλγόριθμο μηχανικής μάθησης:

1. Επιλέγουμε εκ των προτέρων ένα κατάλληλο χώρο υποθέσεων H.

Το απλούστερο παράδειγμα χώρου υποθέσεων είναι το σύνολο των γραμμικών συναρτήσεων

$$\mathcal{H} = \left\{ f : \mathbb{R}^d \to \mathbb{R} : \exists w \in \mathbb{R}^d \text{ such that } f(x) = x^T w, \forall x \in \mathbb{R}^d \right\}$$

2. Μαθαίνουμε την συνάρτηση f (προσδιορίζουμε τις παραμέτρους της) ελαχιστοποιώντας το εμπειρικό σφάλμα στο χώρο υποθέσεων που επιλέξαμε.

$$\hat{f} = \min_{f \in \mathcal{H}} \hat{R}(f)$$

ERM σε σύνθετους χώρους υποθέσεων

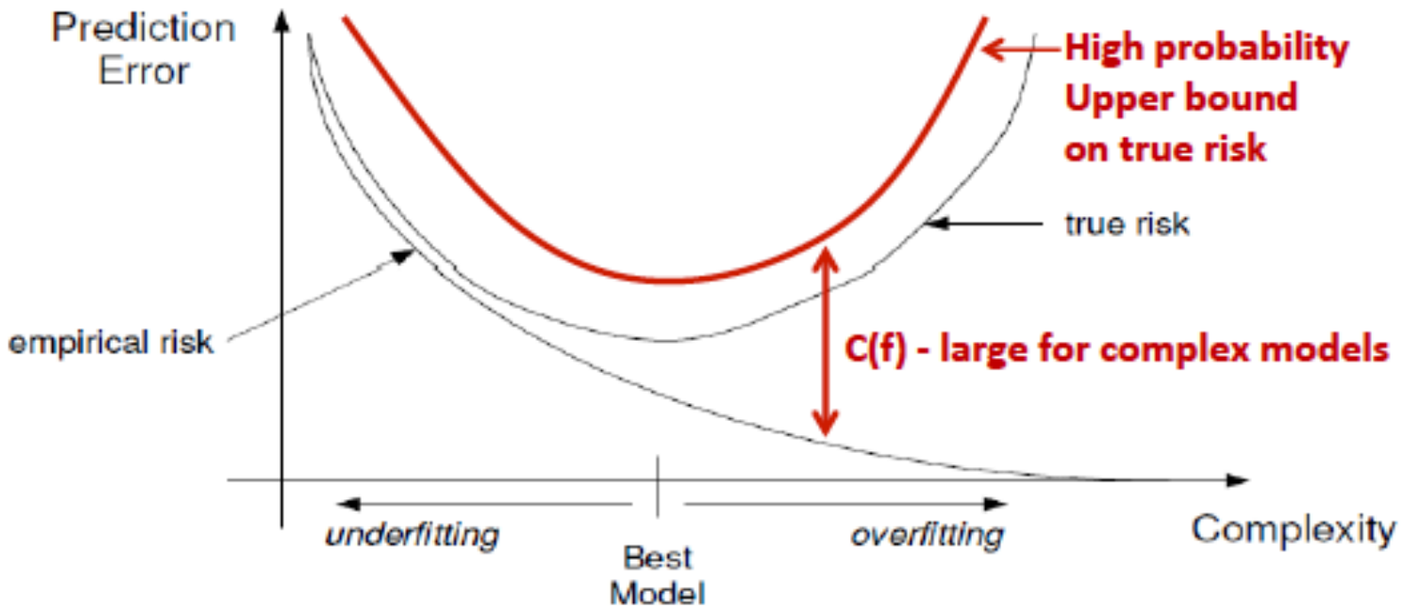
- Εάν ο χώρος υποθέσεων είναι αρκετά πλούσιος (π.χ. το σύνολο των πολυώνυμων βαθμού M>2 ή νευρωνικά δίκτυα μεγάλου βάθους), η επίλυση του ERM συχνά οδηγεί σε overfitting και συνεπώς μεγάλο σφάλμα γενίκευσης.

Ε: Πώς μπορούμε να περιορίσουμε τη διακύμανση της εκτίμησης και να περιορίσουμε το σφάλμα γενίκευσης;

- Οι τεχνικές κανονικοποίησης (regularization techniques) επιτρέπουν ευσταθείς (stable) εκτιμήσεις και τη μείωση του σφάλματος γενίκευσης. Επιπλέον κάποιες τεχνικές κανονικοποίησης επιτρέπουν και μείωση διαστάσεων και συνεπώς συνδράμουν στο περιορισμό της κατάρας των μεγάλων διαστάσεων (curse of dimensionality).

Κανονικοποίηση (Regularization) της πολυπλοκότητας

- Γενική ιδεα structural risk minimization (SRM): Εξισορρόπηση της πολυπλοκότητα του μοντέλου επιβάλοντας "ποινή" στα μοντέλα που αποκλήνουν από το πραγματικό ρίσκο (ανω φράγμα).
- Ισχύει (concentration bounds): $|R(f) - \hat{R}(f)| \leq C(f), \forall f \in \mathscr{H}$



- Structural risk minimization (SRM)

$$\hat{f} = \min_{f \in \mathscr{H}} \{\hat{R}(f) + \lambda C(f)\} = \min_{f \in \mathscr{H}} \{\frac{1}{N} \sum_{i=1}^{N} \ell\left(y_i, f(\mathbf{x}_i)\right) + \lambda C(f)\}$$

- Η συνάρτηση C( ) καλείται regularizer ενώ το λ αποτελεί υπερ-παράμετρο.

SRM και το ξυράφι του Όκαμ (Occam's razor)

- William of Ockham (1285 - 1349) Principles of Parsimony
  Αρχή της Οικονομίας ή Αρχή της Απλότητας

- Ξυράφι του Όκαμ:

«Κανείς δεν θα πρέπει να προβαίνει σε περισσότερες εικασίες από όσες είναι απαραίτητες».

Εναλλακτικά *«Όταν πρέπει να επιλεγεί ένα από δύο μοντέλα με ταυτόσημες προβλέψεις, επιλέγεται το απλούστερο».*

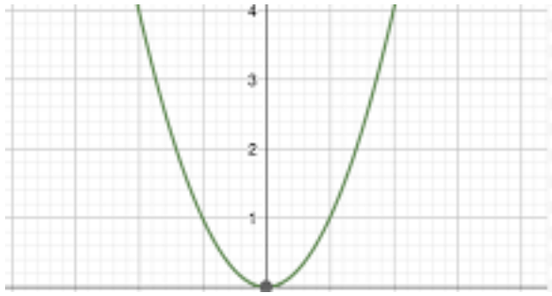
Επιλογή συναρτήσεων κανονικοποίησης

- Η επιλογή διαφορετικών συναρτήσεων κανονικοποιησης (regularizers) και κόστους (loss functions) στο SRM οδηγεί στο σχεδιάσμό διαφορετικών μεθόδων μηχανικής μάθησης, οι οποίες παρουσιάζουν διαφορετικές ιδιότητες.

- Οι πιο συνηθισμένες συναρτήσεις κανονικοποιήσης (regularizers):

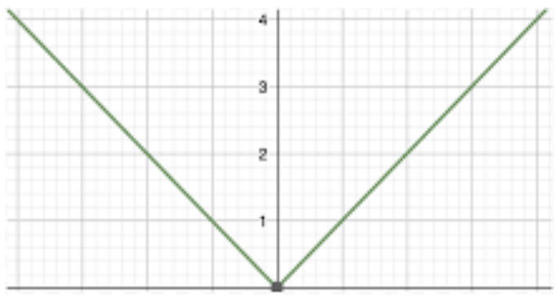
ℓ_2 – norm :
(ή Ευκλείδεια νόρμα)

$$\|\mathbf{x}\|_2^2 = \sum_{i=1}^{d} x_i^2$$



ℓ_1 – norm :

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{d} |x_i|$$



Κανονικοποίηση Tikhonov (Tikhonov regularization)

Κανονικοποίηση Tikhonov

- Το σχήμα κανονικοποίσης του Tikhonov, χρησιμοποιεί την Ευκλείδεια απόσταση ως regularizer στο φορμαλισμό του SRM για την μάθηση παραμετρικών συναρτήσεων των οποίων οι παράμετροι έχουν ελάχιστο μήκος (με την Ευκλείδεια έννοια).

$$\mathbf{w}^* = \min_{\mathbf{w}} \{ \frac{1}{N} \sum_{i=1}^{N} \ell \left(y_i, f(x_i; \mathbf{w}) \right) + \lambda \|\mathbf{w}\|_2^2 \}$$

- Τα μοντέλα (παραμετρικές συναρτήσεις) που προκύπτουν ως λύση του παραπάνω προβλήματος είναι γνωστά και ως δίκτυα κανονικοποίησης (regularization networks).

- Η Ευκλείδια νόρμα αποτελεί τον regularizer στο σχήμα Tikhonov και ελέγχει την ευστάθεια της λύσης εμποδίζοντας το φαινόμενο overfitting.

- Η υπερπαράμετρος λ ισσοροπεί το εμπειρικό σφάλμα και τον regulariser και προσδιορίζεται μέσω cross-validation

Επιλογή συναρτήσεων κόστους στη κανονικοποίηση Tikhonov

$$\mathbf{w}^* = \min_{\mathbf{w}} \{ \frac{1}{N} \sum_{i=1}^{N} \ell \left(y_i, f(x_i; \mathbf{w}) \right) + \lambda \|\mathbf{w}\|_2^2 \}$$

- Διαφορετικές επιλογές της συνάρτησης κόστους (loss functions) οδηγούν σε διαφορετικές μεθόδους μάθησης.

- Δεν υπάρχει γενικός αλγόριθμος επίλυσης του παραπάνω προβλήματος. Η λύση εξαρτάται από την επιλογή της συνάρτησης κόστους.

- Στη συνέχεια αυτού του μαθήματος θα θεωρήσουμε ως συνάρτηση κόστους την τετραγωνική απόσταση:

$$\ell(y, f(x; w)) = (y - f(x; w))^2$$

- Στη περίπτωση αυτή η μέθοδος που προκύπτει ονομάζεται ρυθμισμένα ελάχιστα τετράγωνα (regularized least squares).

Ρυθμισμένα ελάχιστα τετράγωνα

- Η μέθοδος των ρυθμισμένων ελαχίστων τετραγώνων επιδιώκει την εύρεση των άγνωστων παραμέτρων **w** έτσι ώστε να ελαχιστοποιείται η τετραγωνική διαφορά των προβλέψεων από τις πραγματικές μεταβλητές στόχου και επιπλέον το διάνυσμα των παραμέτρων **w** να έχει ελάχιστη Ευκλείδεια νόρμα (μήκος).

$$\mathbf{w}^* = \min_{\mathbf{w}} \{ \frac{1}{N} \sum_{i=1}^{N} \left(y_i - f(x_i; \mathbf{w}) \right)^2 + \lambda \|\mathbf{w}\|_2^2 \}$$

- Εάν επιπλέον η παραμετρική συνάρτηση είναι γραμμική (ως προς τις άγνωστες παραμέτρους) τότε καταλήγουμε στη μέθοδο η οποία είναι γνωστή ως ridge regression:

$$\mathbf{w}^* = \min_{\mathbf{w}} \{ \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T x_i \right)^2 + \lambda \|\mathbf{w}\|_2^2 \}$$

- Τα δεδομένα εισόδου μπορούν να υποστούν κάποιο μη-γραμμικό μετασχηματισμό ϕ():

$$\mathbf{w}^* = \min_{\mathbf{w}} \{ \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T \phi_i(x_i) \right)^2 + \lambda \|\mathbf{w}\|_2^2 \}$$

Ridge regression

- Το πρόβλημα Ridge regression σε μορφή πινάκων εκφράζεται ως εξής:

$$\mathbf{w}^* = \min_{\mathbf{w}}\{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \mathbf{w}^T x_i)\right)^2 + \lambda\|\mathbf{w}\|_2^2\} = \min_{\mathbf{w}}\{\frac{1}{N}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2\}$$

- Ο πίνακας **X** έχει διαστάσεις N x d, δηλαδή περιέχει στις γραμμές του τα δεδομένα d διαστάσεων

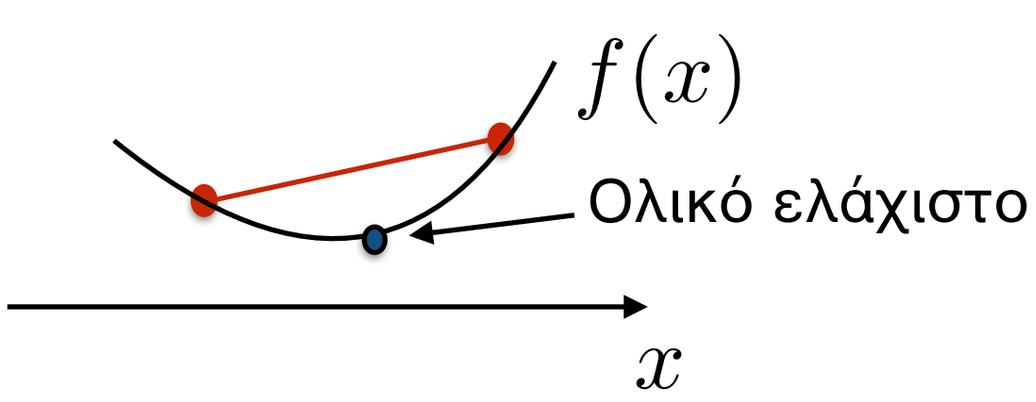
- Το διάνυσμα στήλη **y** έχει διάσταση N και αναπαριστά της μεταβλητές στόχου του συνόλου εκπαίδευσης.

- Το διάνυσμα στήλη **w** έχει διάσταση d και αναπαριστά τις άγνωστες παραμέτρους του μοντέλου.

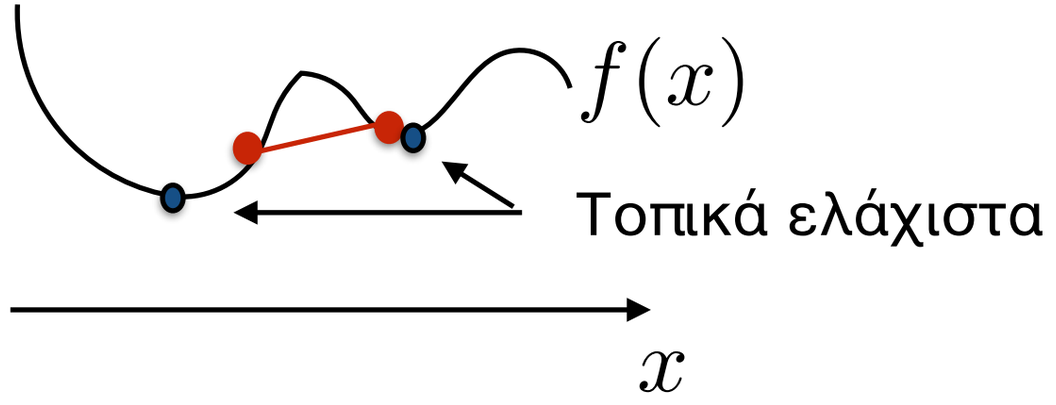
Ε: Πώς βρίσκουμε τη λύση του προβλήματος ridge regression;

Κυρτές συναρτήσεις

Κυρτή (convex) συνάρτηση



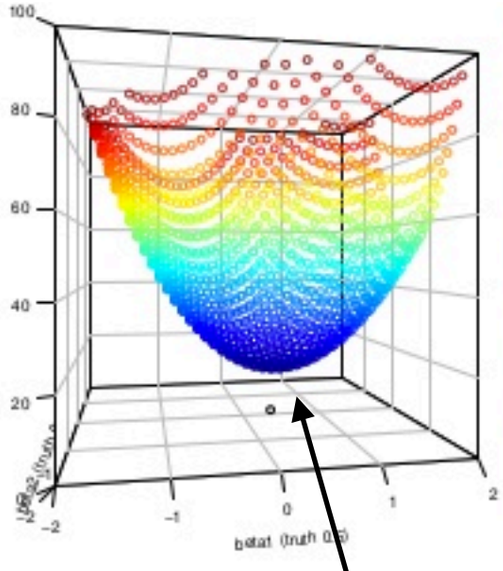
Μη-κυρτή (non-convex) συνάρτηση



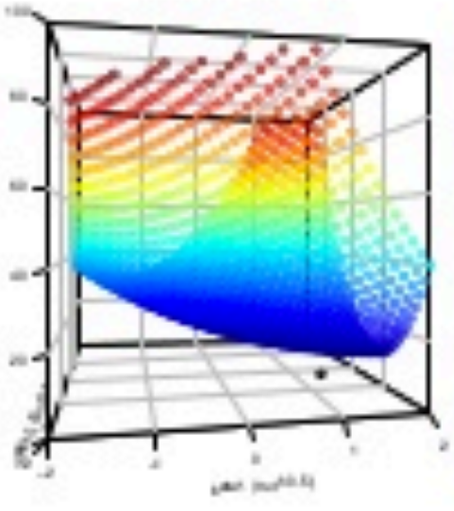
Συνάρτηση ελαχιστοποίησης ridge regression

- Το πρόβλημα Ridge regression απαιτεί να ελαχιστοποιήσουμε ένα άθροισμά Ευκλείδιων αποστάσεων και συνεπώς η συνάρτηση είναι κυρτή με μοναδικό ολικό ελάχιστο.

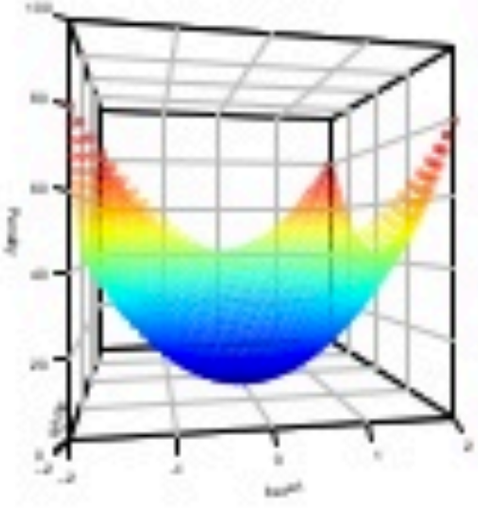
$$\mathbf{w}^* = \min_{\mathbf{w}} l(\mathbf{w}) = \frac{1}{N}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$



$=$



$+$

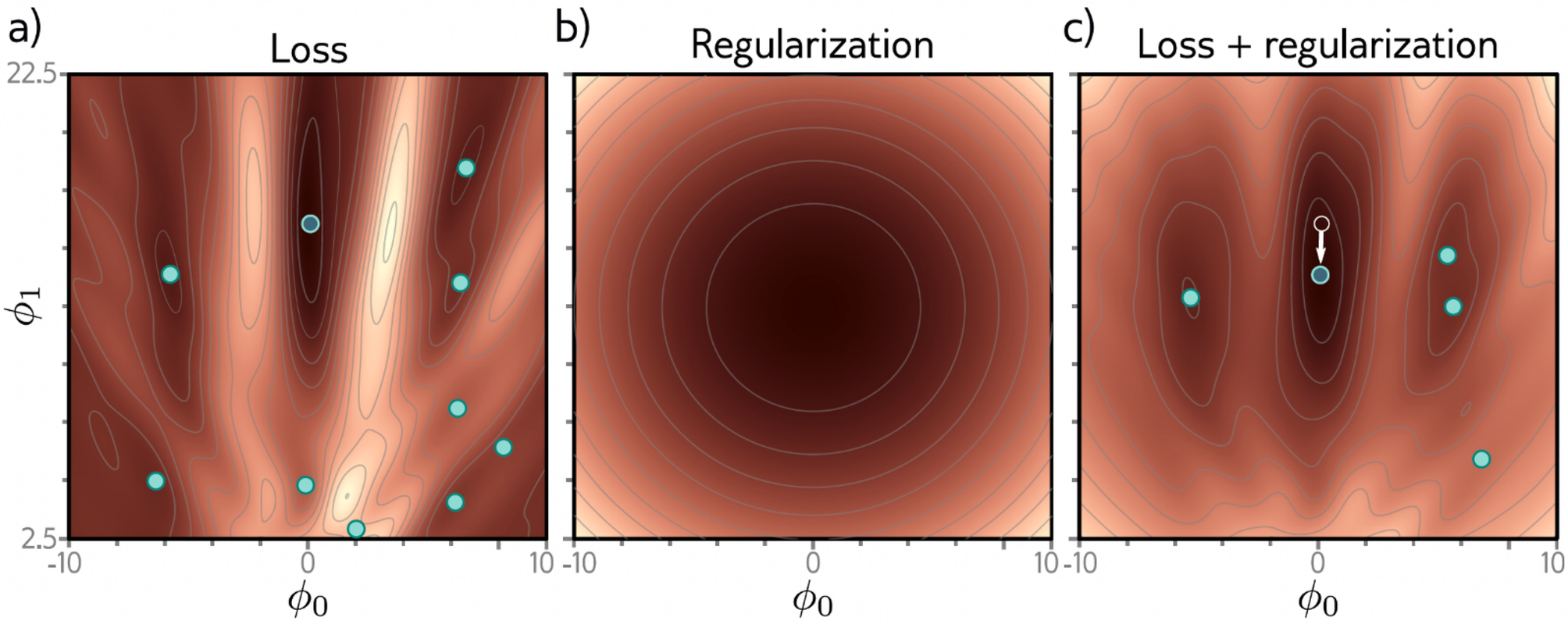


$l(\mathbf{w})$

$\frac{1}{N}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$

$\lambda\|\mathbf{w}\|_2^2$

Ολικό ελάχιστο


a)
Loss
b)
Regularization
c)
Loss + regularization
22.5
2.5
ϕ_1
-10
0
10
ϕ_0

Επίλυση ridge regression

- Εφόσον η συνάρτηση στο πρόβλημα ridge regression είναι συνεχής και παραγωγίσιμη (ώς άθροισμα συνεχών και παραγωγίσιμων συναρτήσεων), η μοναδική λύση του προβλήματος προκύπτει σε κλειστή μορφή μέσω της κανονικής εξίσωσης (normal equation):

1. (Κανονικές εξισώσεις) Υπολογίζουμε τη παράγωγο της συνάρτησης και τη θέτουμε ίση με το μηδενικό διάνυσμα:

$$\nabla_{\mathbf{w}} l(\mathbf{w}) = \mathbf{0}$$

Κανονικές εξισώσεις: $\left(\mathbf{X}^T\mathbf{X} + \lambda N I\right)\mathbf{w} = \mathbf{X}^T\mathbf{y}$

2. Λύνουμε τις κανονικές εξισώσεις:

$$\mathbf{w}^* = \left(\mathbf{X}^T\mathbf{X} + \lambda N I\right)^{-1}\mathbf{X}^T\mathbf{y}$$

Η υπερπαράμετρος λ ελέγχει την αντιστρεψιμότητα του πίνακα: $\mathbf{X}^T\mathbf{X} + \lambda N I$

Ελάχιστα τετράγωνα σε πολλές διαστάσεις

- Όταν το πλήθος των δεδομένων εισόδου (N) είναι μεγαλύτερο από τη διάσταση τους (d), δηλαδή ισχύει d<N τότε η λύση των ελαχίστων τετραγώνων είναι μοναδική και υπολογίζεται σε κλειστή μορφή.

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \underbrace{\|\mathbf{y} - \boldsymbol{X}\mathbf{w}\|_2^2}_{l(\mathbf{w})} \qquad \mathbf{w}* = \underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T}_{\text{pseudoinverse}} \quad \mathbf{y}$$

- Πολλές φορές στη πράξη λόγω της διαστατικότητας των δεδομένων και του υψηλού κόστους συλλογής τους ισχύει d>N. Σε αυτή τη περίπτωση υπάρχει πολύ μεγάλη πιθανότητα να εμφανιστεί το φαινόμενο της συγγραμμικότητας (collinearity):

$$\mathbf{x}_i = \sum_{j \in \mathcal{S}} \mathbf{x}_j a_j$$

Εάν έχουμε ισχυρή συγγραμμικότητα τότε ο πίνακας $\mathbf{X}^T\mathbf{X}$ είναι χαμηλής τάξης και συνεπώς ιδιάζων (μη-αντιστρέψιμος), οπότε το σύστημα κανονικών εξισώσεων δεν μπορεί να λυθεί.

$$\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y}$$

Ελάχιστα τετράγωνα σε πολλές διαστάσεις

- Όταν το πλήθος των δεδομένων εισόδου (N) είναι μεγαλύτερο από τη διάσταση τους (d), δηλαδή ισχύει d<N τότε η λύση των ελαχίστων τετραγώνων είναι μοναδική και υπολογίζεται σε κλειστή μορφή.

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \underbrace{\|\mathbf{y} - \boldsymbol{X}\mathbf{w}\|_2^2}_{l(\mathbf{w})} \qquad \mathbf{w}* = \underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T}_{\text{pseudoinverse}} \quad \mathbf{y}$$

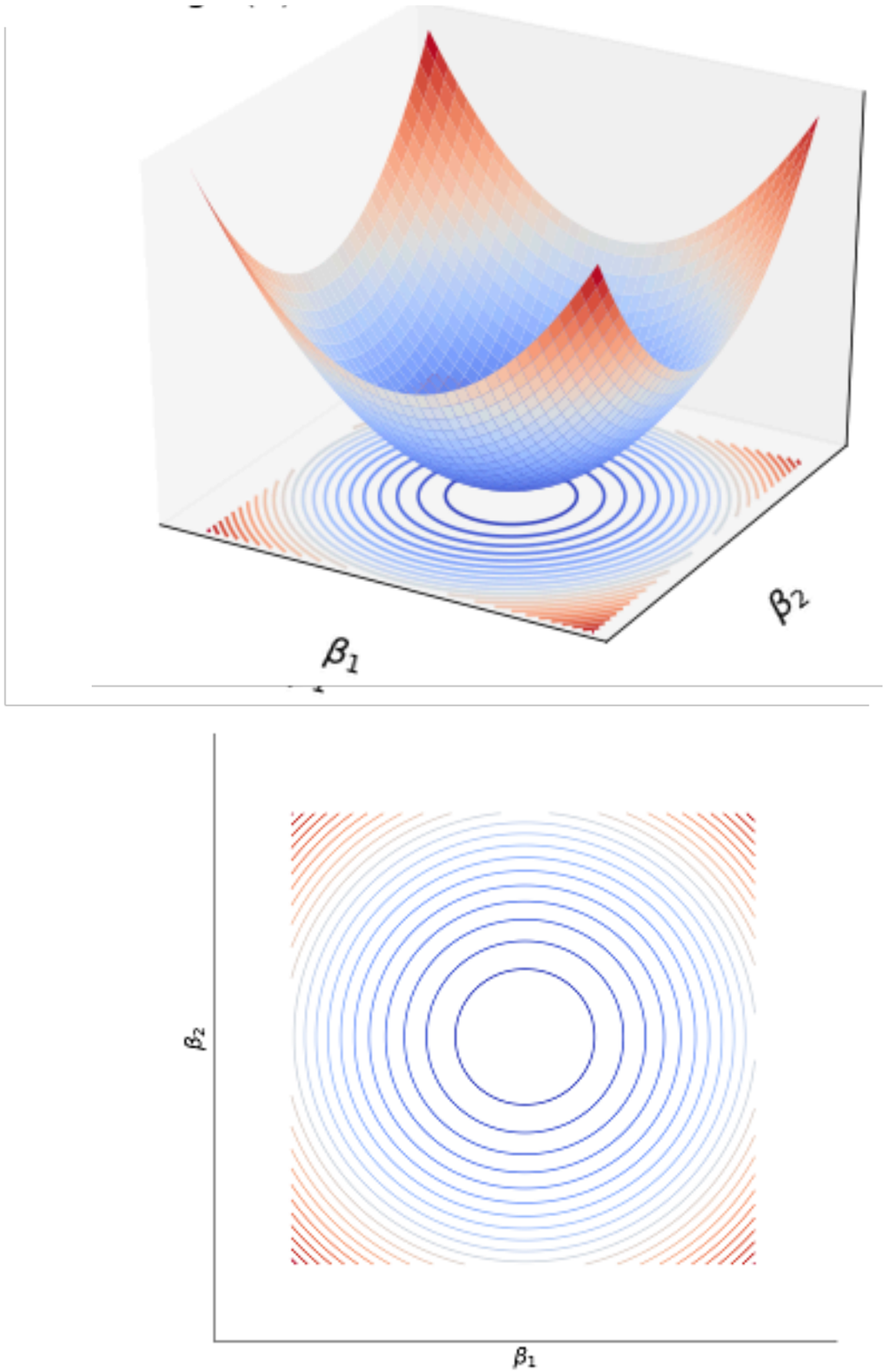
- Πολλές φορές στη πράξη λόγω της διαστατικότητας των δεδομένων και του υψηλού κόστους συλλογής τους ισχύει d>N. Σε αυτή τη περίπτωση υπάρχει πολύ μεγάλη πιθανότητα να εμφανιστεί το φαινόμενο της συγγραμμικότητας (collinearity):

$$\mathbf{x}_i \approx \sum_{j \in \mathcal{S}} \mathbf{x}_j a_j$$

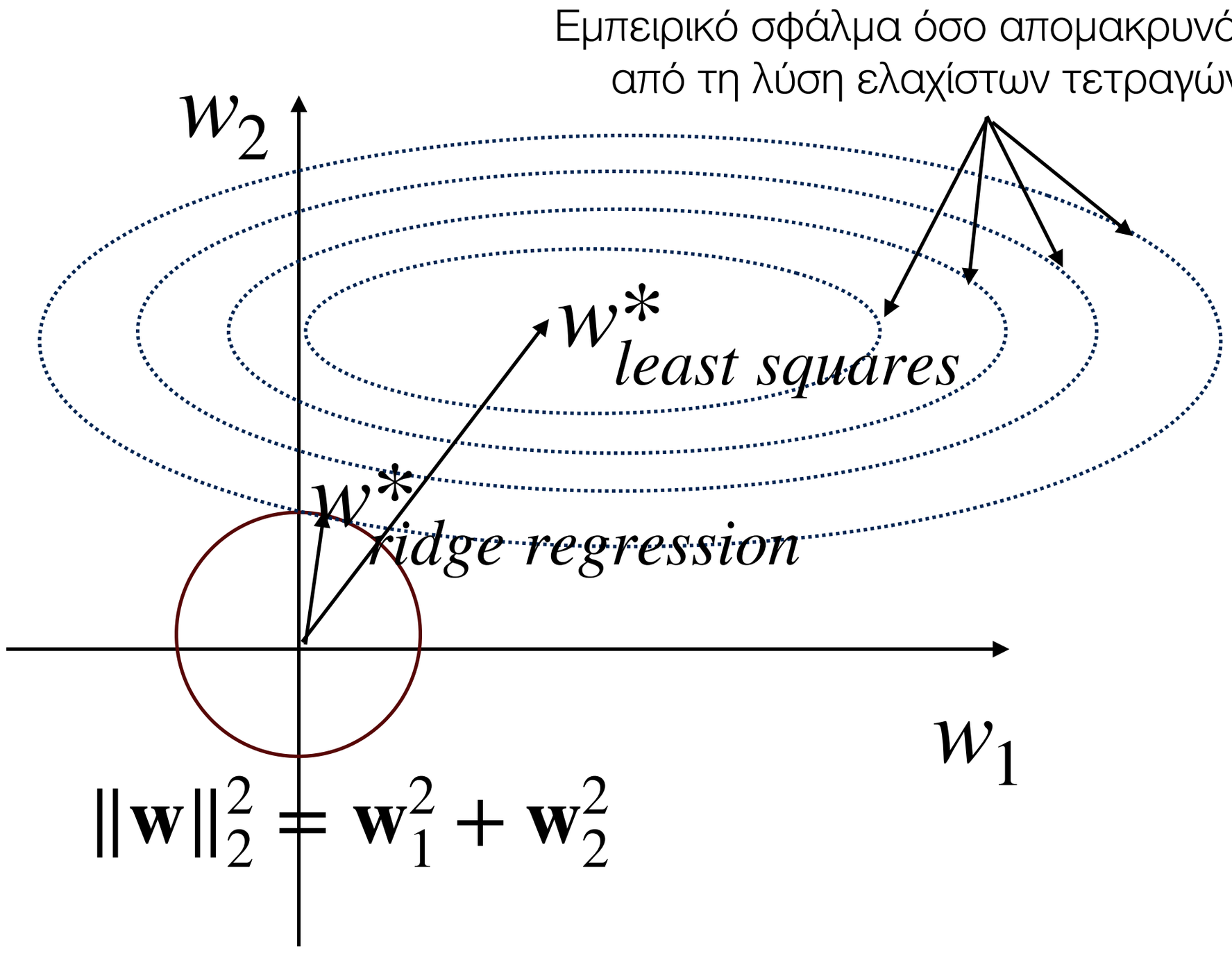
Εάν έχουμε προσεγγιστική συγγραμμικότητα τότε ο $\mathbf{X}^T\mathbf{X}$ είναι κοντά στο να είναι ιδιάζων, ωστόσο αντιστρέψιμος.

- Σε αυτή τη περίπτωση η διακύμανση της εκτίμησης των παραμέτρων είναι μεγάλη και συνεπώς ο εκτιμητής είναι ασταθής (unstable)

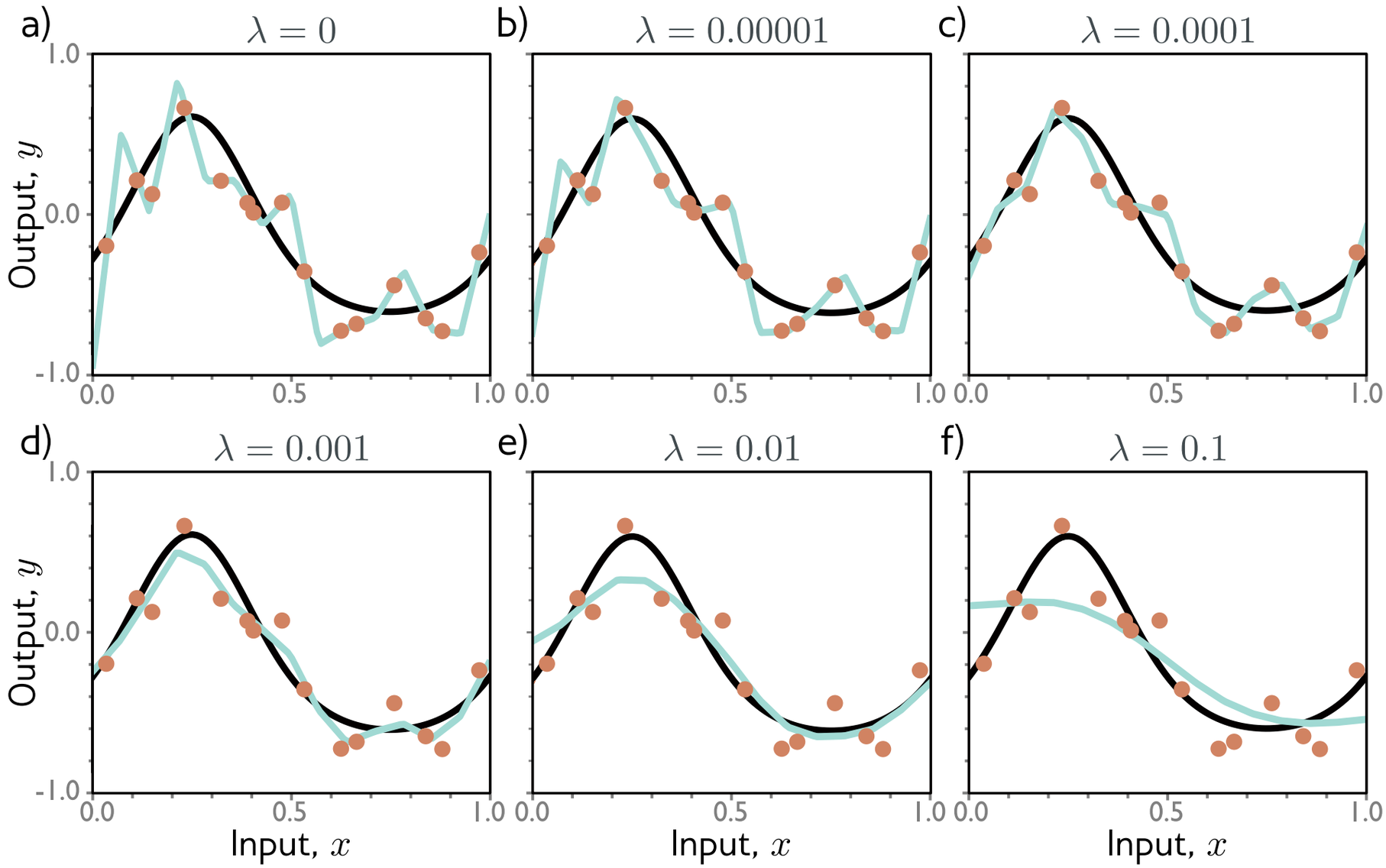
Γεωμετρία λύσεων



Γεωμετρία λύσεων



L2 regularization



Χρησιμότητα της L2 regularization

- **Μειώνει την πολυπλοκότητα του μοντέλου:** Η προσθήκη του όρου κανονικοποίησης L2 στη συνάρτηση απώλειας (loss function) τιμωρεί τα μεγάλα βάρη (weights), οδηγώντας σε ένα απλούστερο μοντέλο με μικρότερα βάρη. Αυτό περιορίζει την ικανότητα του μοντέλου να προσαρμοστεί υπερβολικά στα δεδομένα εκπαίδευσης.

- **Βελτιώνει τη γενίκευση:** Περιορίζοντας το μέγεθος των βαρών, η κανονικοποίηση L2 εμποδίζει το μοντέλο από το να δίνει υπερβολική έμφαση σε μεμονωμένα χαρακτηριστικά (features). Αυτό βοηθά το μοντέλο να γενικεύει καλύτερα σε νέα, άγνωστα δεδομένα.

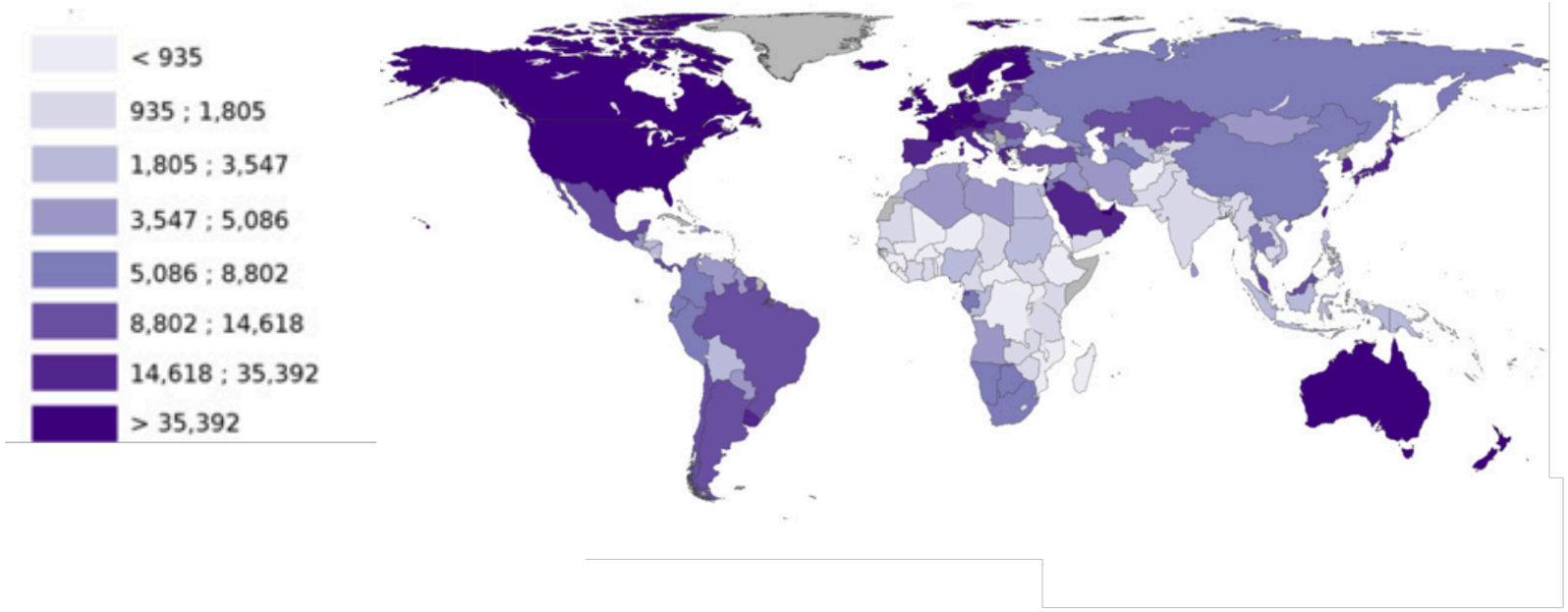
- **Κάνει το μοντέλο πιο ανθεκτικό:** Η κανονικοποίηση L2 κάνει το μοντέλο λιγότερο ευαίσθητο σε μικρές διακυμάνσεις στα δεδομένα εισόδου, καθιστώντας το πιο ανθεκτικό στον θόρυβο και στα ακραία σημεία δεδομένων (outliers).

- **Προωθεί τη διαμοιρασμένη εκμάθηση:** Η κανονικοποίηση L2 ενθαρρύνει το μοντέλο να χρησιμοποιεί όλα τα χαρακτηριστικά (features) αντί να βασίζεται έντονα σε λίγα. Αυτό οδηγεί σε μια πιο ισορροπημένη και διαμοιρασμένη εκμάθηση των χαρακτηριστικών.

Lasso regression, αραιότητα και επιλογή χαρακτηριστικών

Κίνητρο: Επιλογή χαρακτηριστικών μέσω αραιότητας

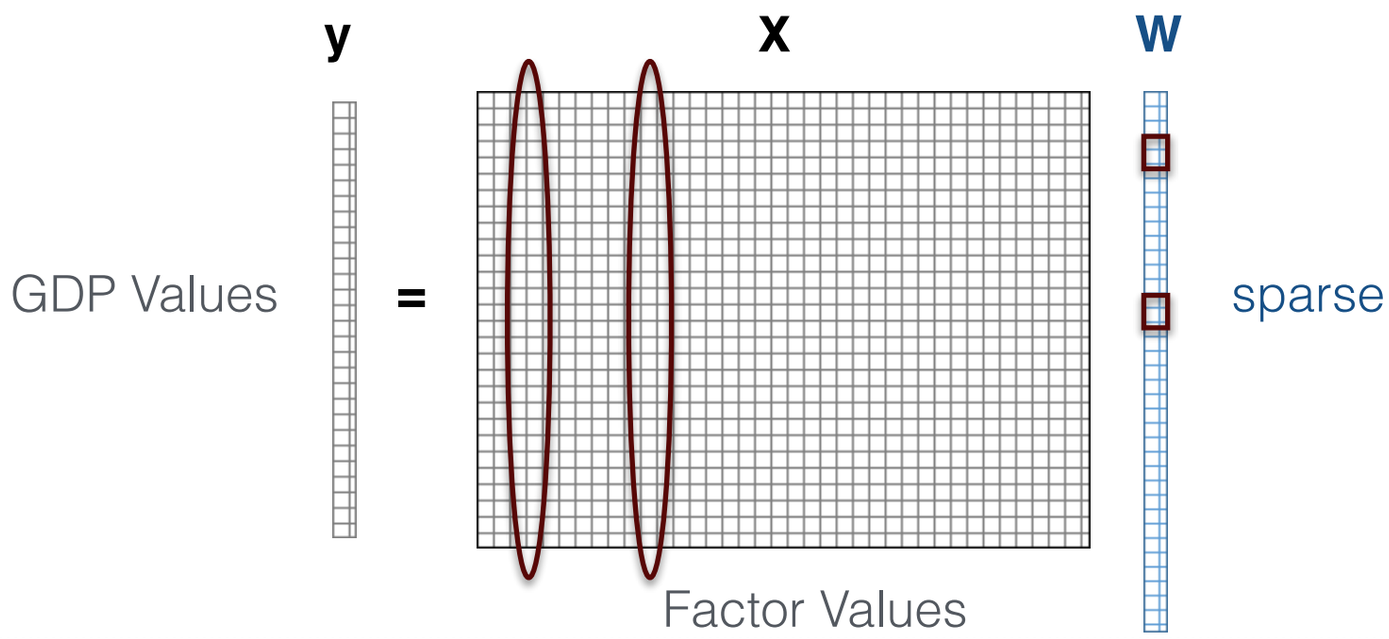
GDP Map



Data - Factors

- Populations density
- Fraction of tropical area
- Size of economy
- Defence spending
- Public investment
- Life expectancy
  - .
  - .
  - .
- Land area

Ποιοι παράγοντες σχετίζονται με το ΑΕΠ κάθε χώρας;



Lasso regression

Ε: Πώς βρίσκουμε αραιές (sparse) παραμέτρους σε προβλήματα παλινδρόμησης;

- Ορισμός: Ένα διάνυσμα είναι αραιό (sparse) όταν τα περισσότερα στοιχεία του είναι μηδενικά και μόνο ένα μικρό πλήθος στοιχείων είναι μή μηδενικό.

- Για να βρούμε αραιά διανύσματα παραμέτρων αρκεί να λύσουμε ένα σταθμισμένο πρόβλημα ελαχίστων τετραγώνων με regularizer την l1-νόρμα. Το πρόβλημα αυτό είναι γνωστό ως Lasso regression και εκφράζεται μαθηματικά ως εξής:

$$\mathbf{w}^* = \min_{\mathbf{w}} \{ \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T x_i \right)^2 + \lambda \|\mathbf{w}\|_1 \} = \min_{\mathbf{w}} \{ \frac{1}{N} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1 \}$$

- Ο πίνακας **X** έχει διαστάσεις N x d, δηλαδή περιέχει στις γραμμές του τα δεδομένα d διαστάσεων
- Το διάνυσμα στήλη **y** έχει διάσταση N και αναπαριστά της μεταβλητές στόχου του συνόλου εκπαίδευσης.
- Το διάνυσμα στήλη **w** έχει διάσταση d και αναπαριστά τις άγνωστες παραμέτρους του μοντέλου, και μόνο ένα μικρό υποσύνολο τους έχουν μη-μηδενικές τιμές.

Γεωμετρία λύσεων Lasso regression και Ridge regression

Fig 8(a) : L1 and L2 Norms

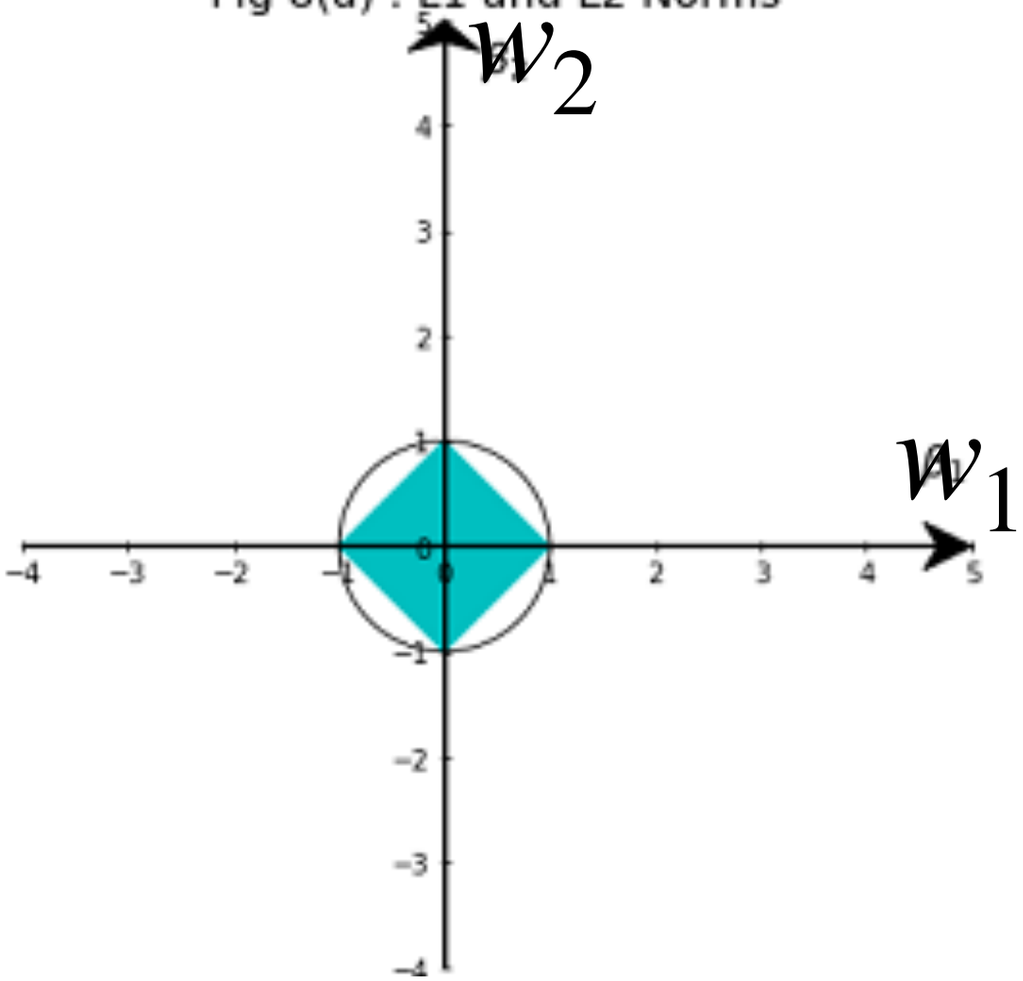
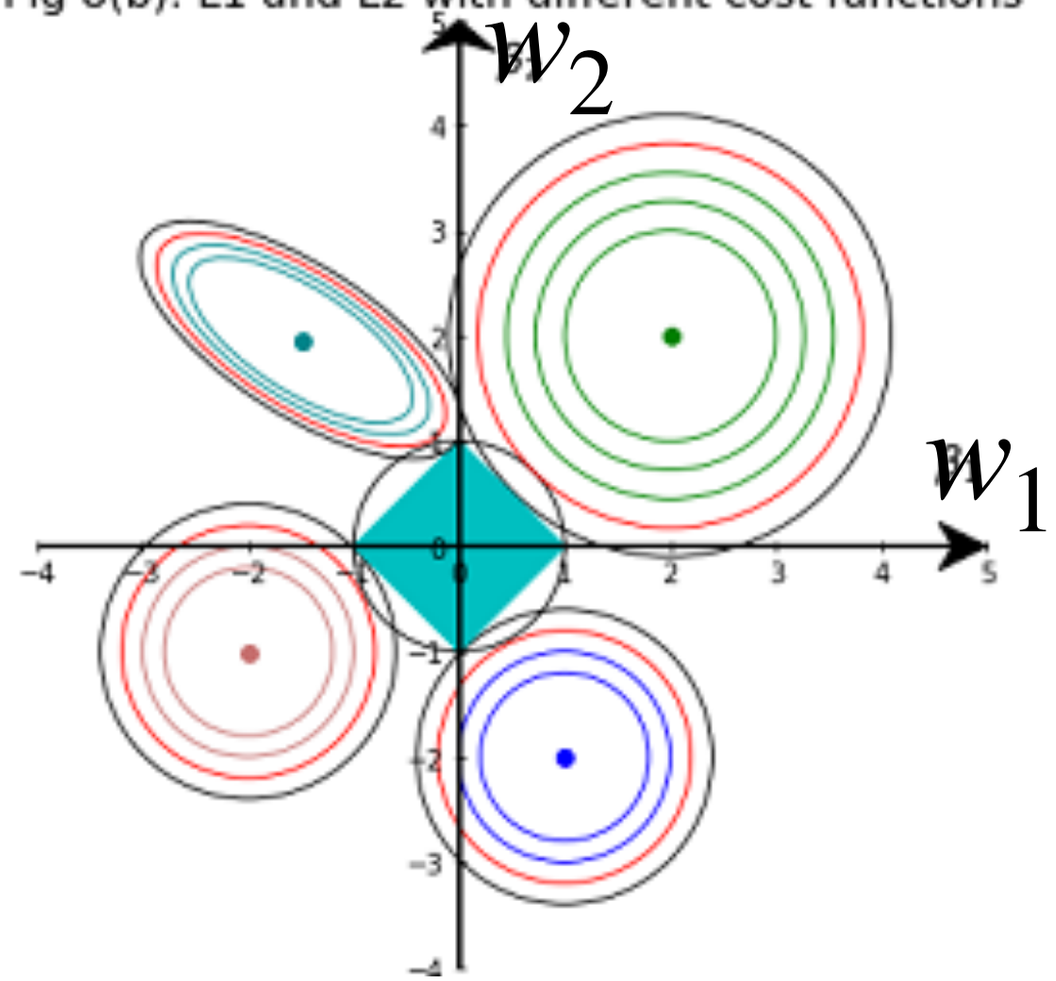


Fig 8(b): L1 and L2 with different cost functions



Χρησιμότητα της L1 regularization

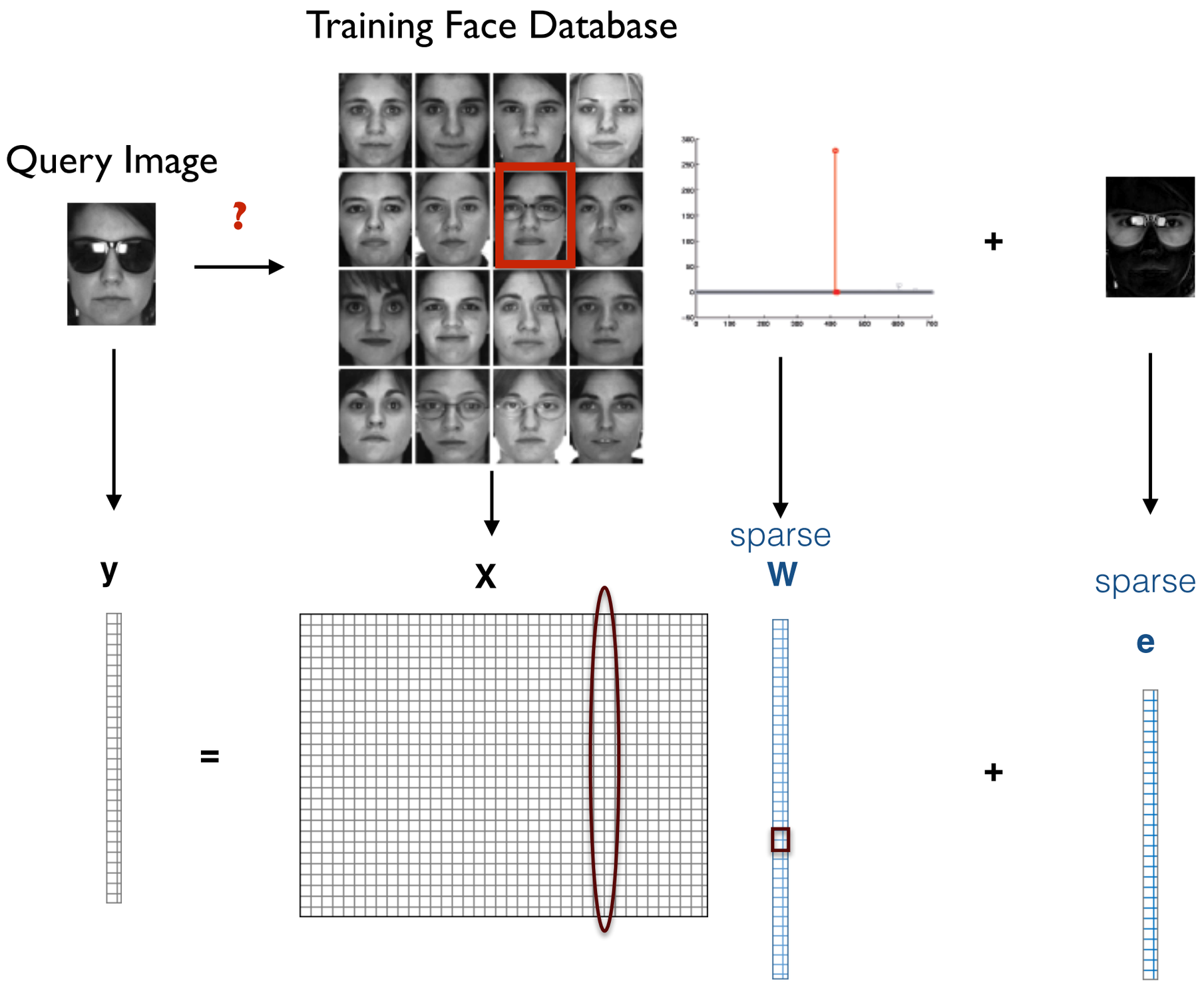
- **Επιλογή χαρακτηριστικών (Feature selection)**: Η κανονικοποίηση L1 έχει την ιδιότητα να οδηγεί μερικά βάρη (weights) του μοντέλου σε ακριβώς μηδέν. Αυτό ουσιαστικά εκτελεί επιλογή χαρακτηριστικών, αφού τα χαρακτηριστικά με μηδενικά βάρη μπορούν να αγνοηθούν. Αυτό είναι ιδιαίτερα χρήσιμο όταν έχουμε ένα μοντέλο με πολλά χαρακτηριστικά, μερικά από τα οποία μπορεί να είναι άσχετα ή περιττά.

- **Ερμηνευσιμότητα του μοντέλου (Model interpretability):** Επειδή η κανονικοποίηση L1 εκτελεί επιλογή χαρακτηριστικών, τα προκύπτοντα μοντέλα τείνουν να είναι πιο αραιά και ευκολότερα ερμηνεύσιμα. Με λιγότερα χαρακτηριστικά στο μοντέλο, είναι ευκολότερο να κατανοήσουμε τη σχέση μεταξύ των εισόδων και των εξόδων.

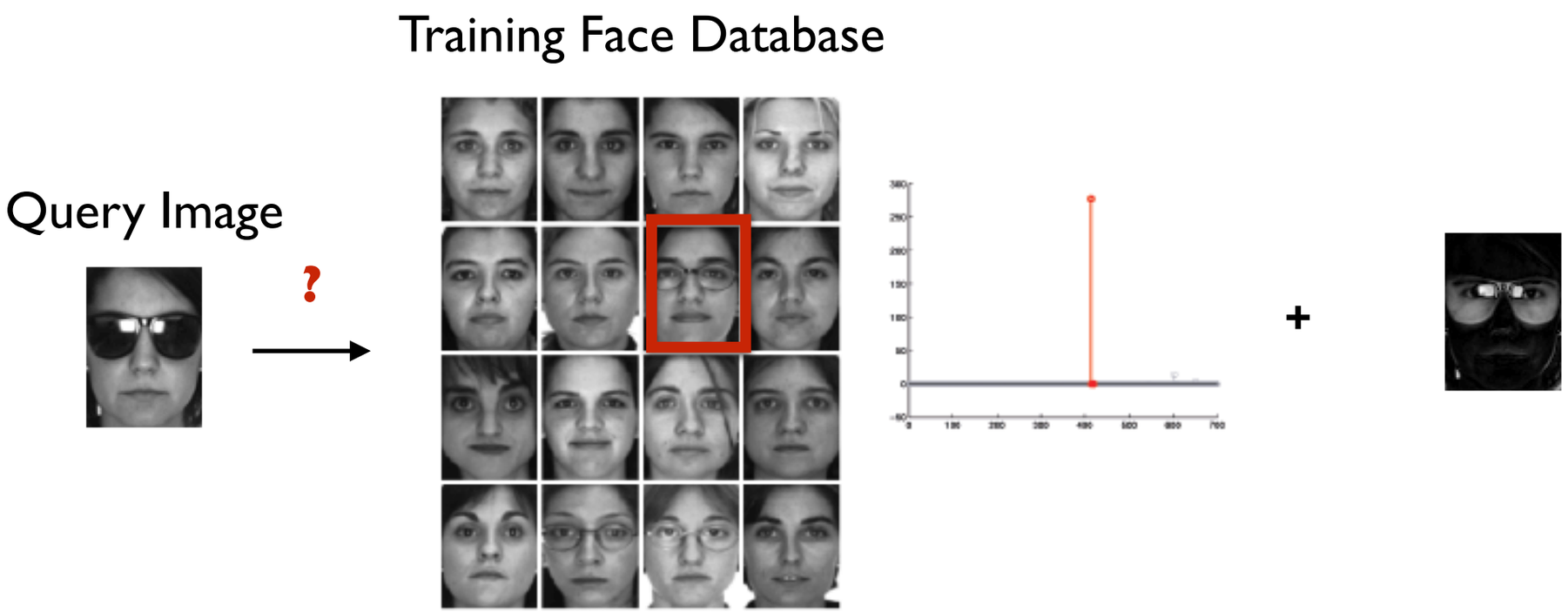
- **Έλεγχος της πολυπλοκότητας του μοντέλου**: Παρόμοια με την κανονικοποίηση L2, η κανονικοποίηση L1 μπορεί να βοηθήσει στον έλεγχο της πολυπλοκότητας του μοντέλου, μειώνοντας την υπερπροσαρμογή (overfitting). Ωστόσο, λόγω της ιδιότητας επιλογής χαρακτηριστικών, η κανονικοποίηση L1 μπορεί συχνά να οδηγήσει σε απλούστερα μοντέλα σε σύγκριση με την κανονικοποίηση L2.

- **Ανθεκτικότητα σε ακραία σημεία δεδομένων (outliers):** Η κανονικοποίηση L1 είναι λιγότερο ευαίσθητη σε ακραία σημεία δεδομένων σε σύγκριση με την κανονικοποίηση L2.

Εύρωστη αναγνώριση προσώπου με Lasso



Εύρωστη αναγνώριση προσώπου με Lasso



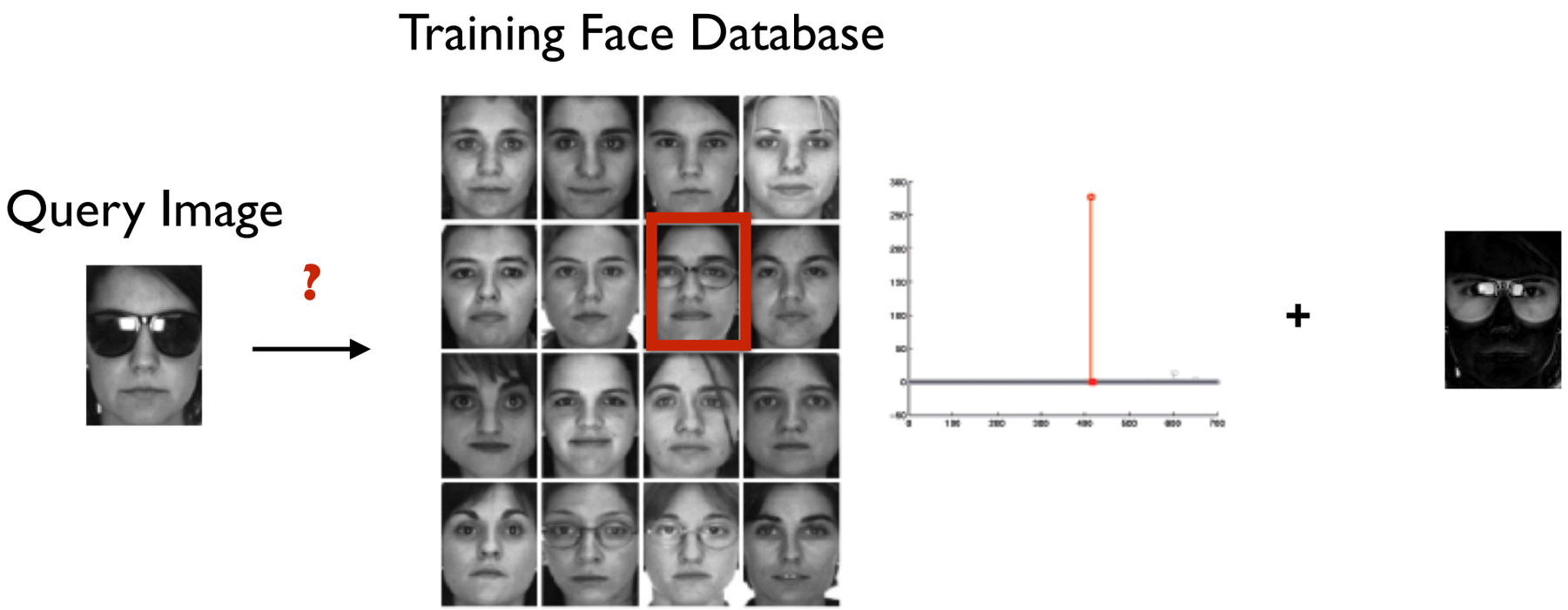
$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}$$

- Το διάνυσμα στήλη **w** έχει διάσταση N και αναπαριστά τους αγνώστους συντελεστές του αραιού γραμμικού συνδυασμού.

- Το διάνυσμα στήλη **e** έχει διάσταση d (πλήθος pixels στην εικόνα) και αναπαριστά τον αραιό θόρυβο (π.χ. occlusions)

$$\mathbf{y} = [\mathbf{X} \;\; \mathbf{I}] \begin{bmatrix} \mathbf{w} \\ \mathbf{e} \end{bmatrix} = \mathbf{A}\mathbf{c}$$

Εύρωστη αναγνώριση προσώπου με Lasso



$$\mathbf{c}^* = \min_{\mathbf{c}} \|\mathbf{y} - \mathbf{X}\mathbf{c}\|_2^2 + \lambda\|\mathbf{c}\|_1$$

Συμπεράσματα

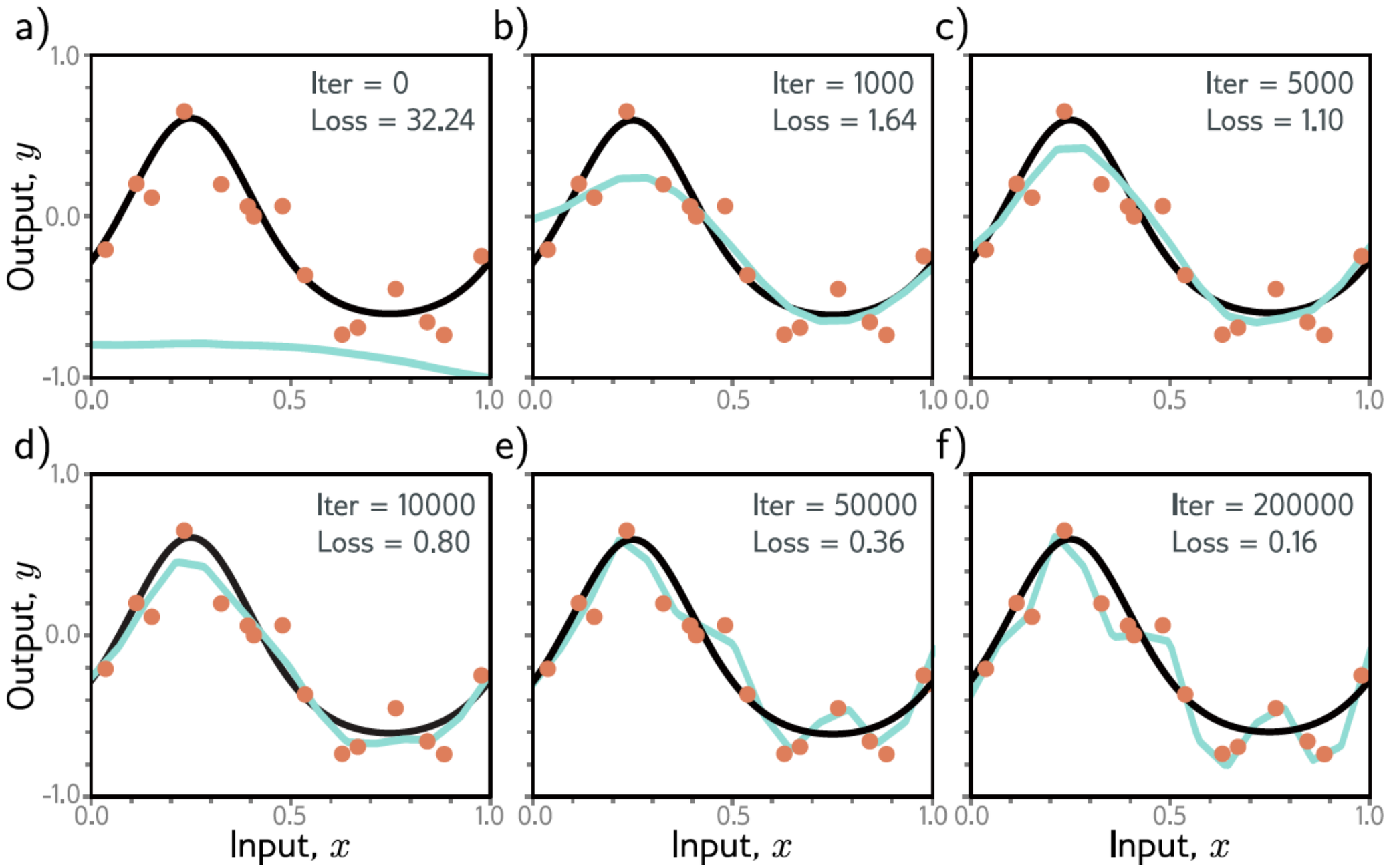
- Παλινδρόμηση ελαχίστων τετραγώνων = πρόβλεψη
- Παλινδρόμηση ελαχίστων τετραγώνων + Tikhonov reg. = πρόβλεψη + ευστάθεια
- Παλινδρόμηση ελαχίστων τετραγώνων + l1 reg.(lasso) = πρόβλεψη + ευστάθεια + επιλογή χαρακτηριστικών (μείωση διαστάσεων)

Έμμεση κανονικοποίηση (implicit regularisation)

Early stopping

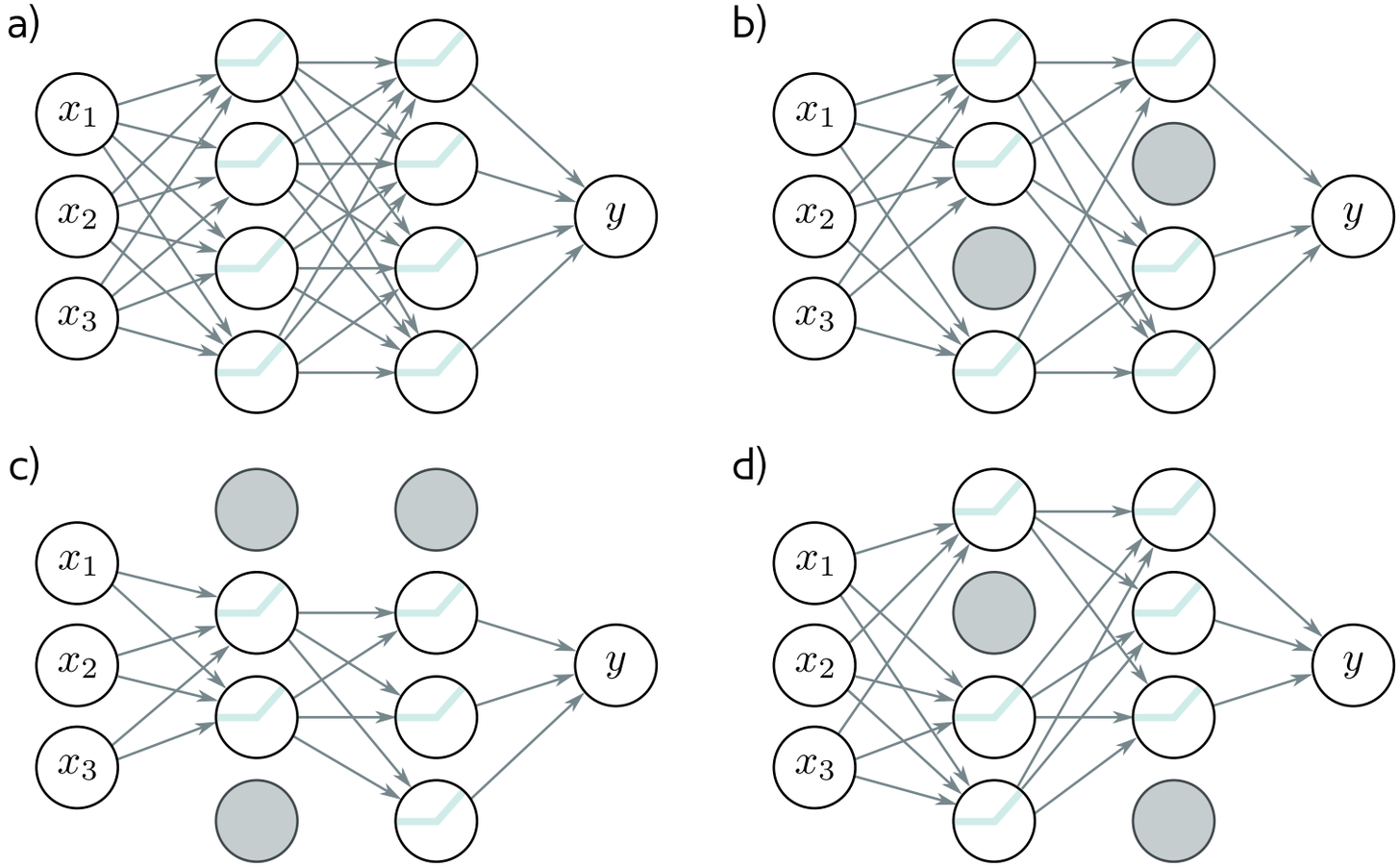
- Αν σταματήσουμε την εκπαίδευση νωρίς, τα βάρη δεν έχουν χρόνο να υπερπροσαρμοστούν στο θόρυβο
- Τα βάρη ξεκινούν μικρά και δεν έχουν χρόνο να μεγαλώσουν
- Μειώνει την πραγματική πολυπλοκότητα του μοντέλου

Early stopping

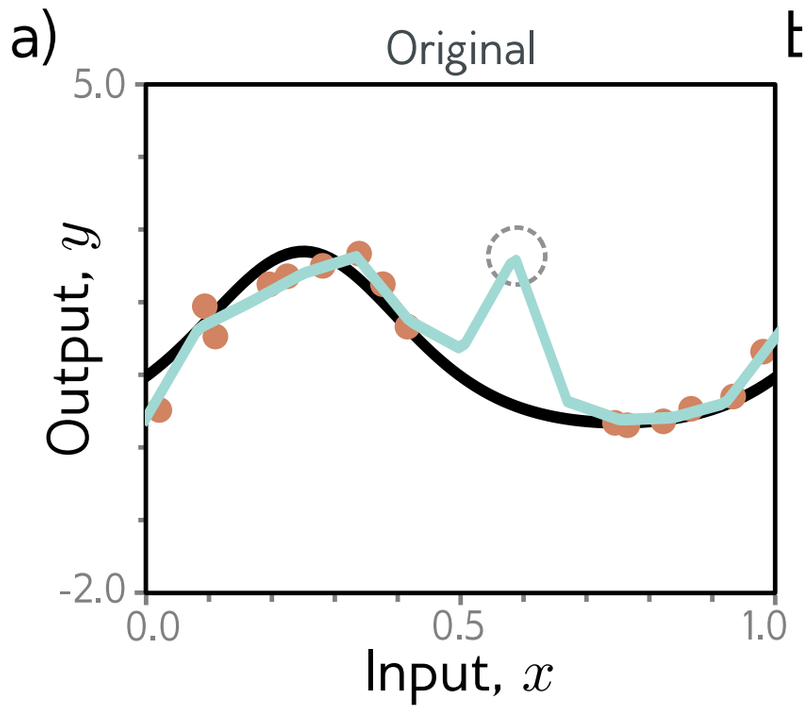


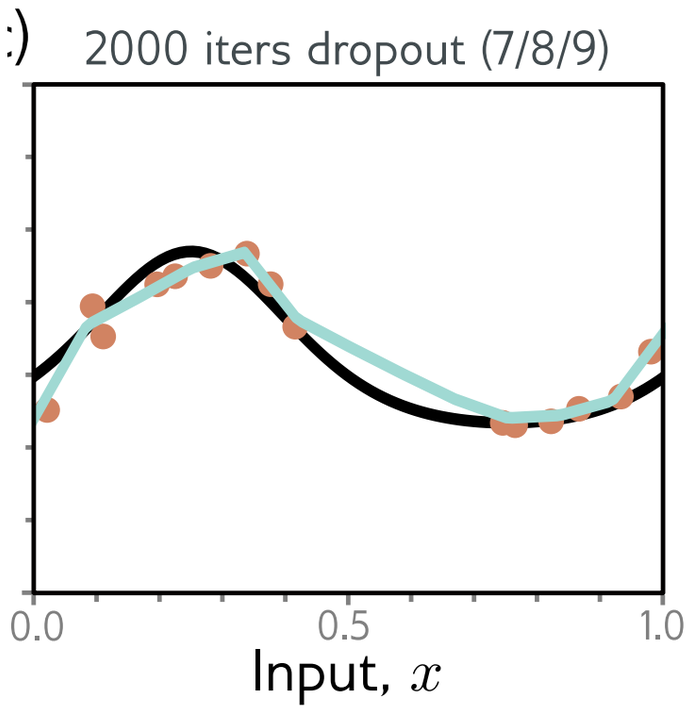
Dropout

- Το Dropout είναι μια τεχνική κανονικοποίησης (regularization) που χρησιμοποιείται συχνά στα νευρωνικά δίκτυα για τη μείωση της υπερπροσαρμογής (overfitting) και τη βελτίωση της γενίκευσης του μοντέλου. Η βασική ιδέα πίσω από το Dropout είναι η τυχαία απενεργοποίηση (ή "απόρριψη") ορισμένων νευρώνων σε ένα νευρωνικό δίκτυο κατά τη διάρκεια της εκπαίδευσης.

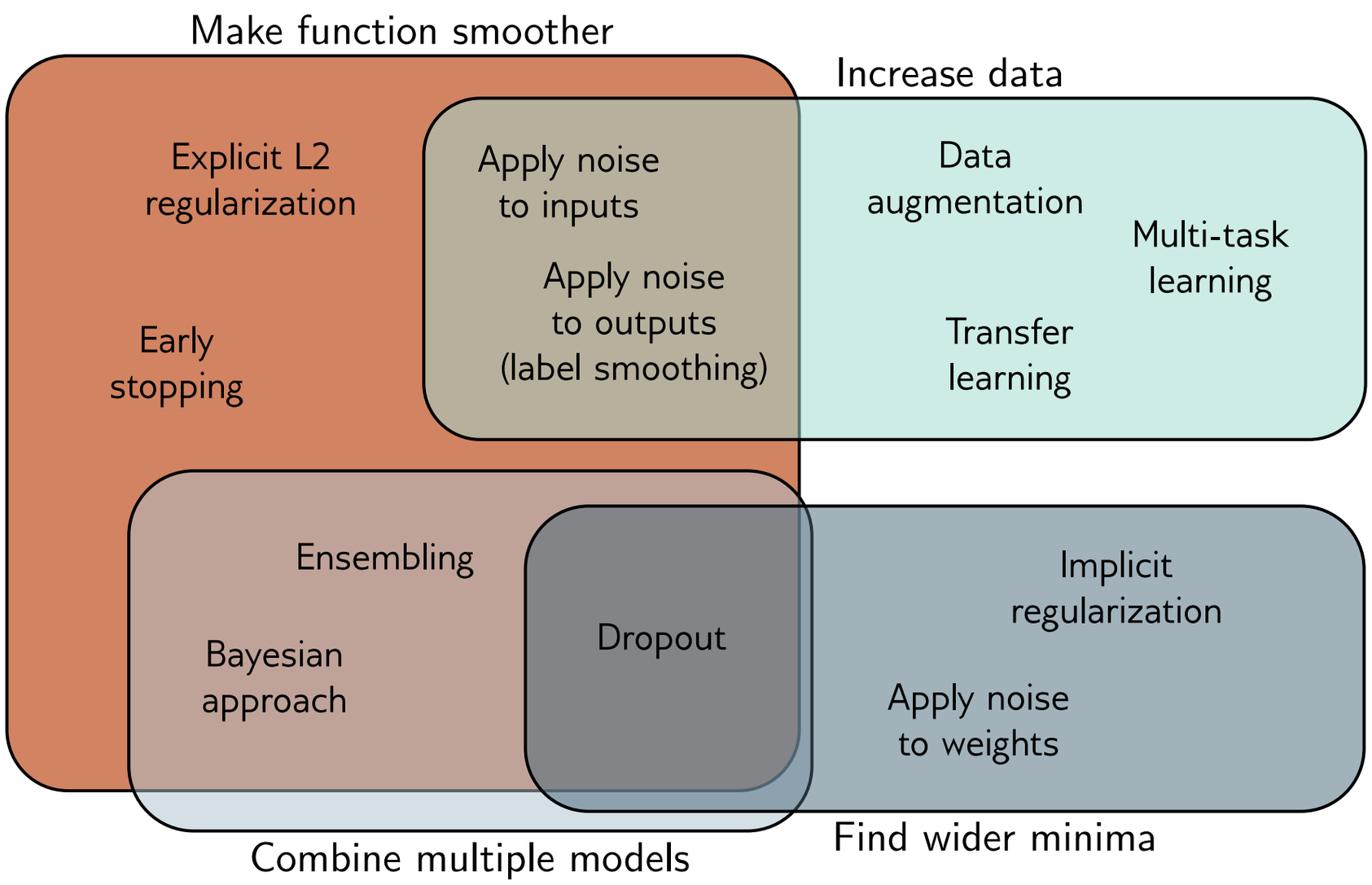


Dropout


a)
Original
5.0
-2.0
0.0
0.5
1.0
Output, y
Input, x


2000 iters dropout (7/8/9)
0.0
0.5
1.0
Input, x

Overview


Make function smoother
Increase data
Explicit L2 regularization
Early stopping
Apply noise to inputs
Apply noise to outputs (label smoothing)
Data augmentation
Multi-task learning
Transfer learning
Ensembling
Bayesian approach
Dropout
Implicit regularization
Apply noise to weights
Combine multiple models
Find wider minima