Topics in Learning Theory

Lecture 5: Regularization

Topics

- Linear classification and regularization
- Rademacher complexity analysis for linear regularization
- L_∞ Covering number for linear regularization
- Regularization and stability

Linear Classifier

- $f(x) = w^T x$, where $x \in R^d$
- classificaiton rule: $y = sign(w^T x)$
- VC theory: without restriction, the complexity term is $O(d \ln n/n)$ (realizable case) or $O(\sqrt{d/n})$ (unrealizable case)
- Can we do better? under margin condition?
  - better estimation of L_∞ covering or rademacher complexity
  - key: complexity independent (or weakly dependent) of d
  - works on modern datasets with large dimensionality.

Regularization conditions

- Restrict the size of w: put additional constraint

$$g(w) \leq a$$

- Example regularization conditions:
  - 2-norm $g(w) = \|w\|_2$
  - L_0: $g(w) = \|w\|_0 = |\{j : w_j \neq 0\}|$ (sparsity)
  - 1-norm $g(w) = \|w\|_1$ (approximate sparsity)
  - L_p: $g(w) = \|w\|_p$
  - entropy: $w_j \geq 0$, $\sum_j w_j = 1$, and $g(w) = \sum_j w_j \ln w_j/\mu_j$, where $\sum_j \mu_j = 1$ ($\mu_j \geq 0$)

Covering number bounds for regularized linear classifiers

- How to measure the complexity of regularized linear function $f(x) = w^T x$: $g(w) \leq a$?

- Bound empirical L_∞-covering number with q-norm regularization

- $p - q$ norm regularization

  If $\|x\|_p \leq b$ and $\|w\|_q \leq a$, where $2 \leq p < \infty$ and $1/p + 1/q = 1$, then $\forall \epsilon > 0$,

$$\ln N_\infty(\mathcal{H}, \epsilon, n) \leq 36(p-1)\frac{a^2 b^2}{\epsilon^2} \ln[2\lceil 4ab/\epsilon + 2 \rceil n + 1].$$

  – independent of dimensionality

- Entropy regularization

  Given μ such that $\sum_j \mu_j = 1$ ($\mu_j \geq 0$) if $\|x\|_\infty \leq b$ and $\|w\|_1 \leq a$ and $\sum_j w_j \ln \frac{w_j}{\mu_j \|w\|_1} \leq c$ ($w_j \geq 0$), then $\forall \epsilon > 0$,

$$\ln \mathcal{N}_\infty(\mathcal{H}, \epsilon, n) \leq \frac{36b^2(a^2 + ac)}{\epsilon^2} \ln[2\lceil 4ab/\epsilon + 2\rceil n + 1].$$

- L_1 regularization: $\|x\|_\infty \leq b$ and $\|w\|_1 \leq a$

  take $\mu_j = 1/d$, then entropy is upper bounded by $\|w\|_1 \ln d$, thus can take $c = a \ln d$:

$$\ln \mathcal{N}_\infty(\mathcal{H}, \epsilon, n) \leq \frac{36b^2a^2(1 + \ln d)}{\epsilon^2} \ln[2\lceil 4ab/\epsilon + 2\rceil n + 1].$$

  – $\ln d$ dependency — weak dependency on dimensionality

L_∞-cover Margin bound

- Consider normalized 2-norm regularization
  - $\|x\|_2 \leq 1$
  - $\|w\|_2 \leq 1$
- Given any fixed λ with probability $1-\eta$, we have the following bound for all $f \in \mathcal{H}$ and all $\gamma \in (0,1]$:

$$\mathbf{E}_{X,Y} I(f(X)Y \leq 0) \leq \frac{1}{(1-\alpha)n} \sum_{i=1}^{n} I(f(X_i)Y_i \leq \gamma) + C \frac{\ln(n/\eta) + \ln(1/\gamma)}{\lambda(1-\alpha)n\gamma^2},$$

where $\lambda = 2(e^\lambda - \lambda - 1)/\lambda$.

Classification-error $\leq$ const * margin-error + $O(\ln n/n)$

- For 1-norm: a similar bound holds: $\|w\|_1 \leq 1$ and $\|x\|_\infty \leq 1$

  Classification-error $\leq$ const * margin-error + $O(\ln d \ln n/n)$

- If the data is dense, with $\|x\|_\infty \leq 1$, $\|x\|_2$ can be as large as $\sqrt{d}$.
  - for dense data, 1-norm regularization has weaker dependency on dimensionality ($\ln d$) than 2-norm regularization (d)

Rademacher Complexity bounds for regularized linear classifiers

- Assume $\|x\|_p \leq a$ and $\|w\|_q \leq b$, where $p \in [2, \infty)$ and $1/p + 1/q = 1$, then

$$R(\mathcal{H}, S_n) \leq \frac{\sqrt{p-1}ab}{\sqrt{n}}.$$

  where $\mathcal{H} = \{f(x) = w^T x; \|x\|_p \leq a, \|w\|_q \leq b\}$.

- Similar result holds for entropy/L_1 regularization.

Proof

Recall $\sigma_i = \pm 1$ with probability 0.5, and

$$
\begin{aligned}
R(S_n) =& E_\sigma \sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) = E_\sigma \sup_{\|w\|_q \leq b} \frac{w^T}{n} \sum_{i=1}^n \sigma_i X_i \\
\leq & b E_\sigma \| \frac{1}{n} \sum_{i=1}^n \sigma_i X_i \|_p \leq \frac{b}{n} (E_\sigma \| \sum_{i=1}^n \sigma_i X_i \|_p^2)^{1/2}
\end{aligned}
$$

Now, we only need to prove that

$$
E_\sigma \| \sum_{i=1}^n \sigma_i X_i \|_p^2 \leq (p-1) \sum_{i=1}^n \|X_i\|_p^2 .
$$

To show this, we let $f(x) = \|x\|_p^2$, and note that $d^2 f(x+tx')/dt^2 \leq 2(p-1)\|x'\|_p^2$. Using Taylor expansion:

$$
\begin{aligned}
&E_\sigma \| \sum_{i=1}^{n} \sigma_i X_i \|_p^2 = E_\sigma \frac{f(\sum_{i=1}^{n-1} \sigma_i X_i + X_n) + f(\sum_{i=1}^{n-1} \sigma_i X_i - X_n)}{2} \\
=& E_\sigma \| \sum_{i=1}^{n-1} \sigma_i X_i \|_p^2 + E_\sigma \frac{d^2}{dt^2} \frac{f(\sum_{i=1}^{n-1} \sigma_i X_i + tX_n) + f(\sum_{i=1}^{n-1} \sigma_i X_i - tX_n)}{4} \\
\leq& E_\sigma \| \sum_{i=1}^{n-1} \sigma_i X_i \|_p^2 + (p-1) \|X_n\|_p^2 \\
\leq& \cdots \leq (p-1) \sum_{i=1}^{n} \|X_i\|_p^2.
\end{aligned}
$$

Rademacher Process Comparison Theorem

- Let $\phi(f, y)$ be Lipschitz in f with constant γ: $|\phi(f, y) - \phi(f', y)| \leq \gamma|f - f'|$, then
$$R(\phi(\mathcal{H})|S_n) \leq \gamma R(\mathcal{H}|S_n).$$

- Can estimate the Rademacher complexity of $\phi(w^T x, y)$ using an estimate of Rademacher complexity of $w^T x$.

Rademacher Margin bound

Let $\phi(f(x), y) = I(f(x)y \leq 0) + I(0 \leq f(x)y \leq \gamma)(1 - f(x)y/\gamma)$, then ϕ is Lipschitz constant $1/\gamma$.

Assume $\|x\|_p \leq a$ and $\|w\|_q \leq b$, where $q \in [2, \infty]$ and $1/p + 1/q = 1$, then

$$E_{X,Y}\phi(f(X), Y) \leq \frac{1}{n}\sum_{i=1}^{n}\phi(f(X_i), Y_i) + \frac{2\sqrt{p-1}ab}{\gamma\sqrt{n}} + \sqrt{\frac{\ln(1/\eta)}{2n}}.$$

Implying margin bound:

$$E_{X,Y}I(f(X)Y \leq 0) \leq \frac{1}{n}\sum_{i=1}^{n}I(f(X_i)Y_i \leq \gamma) + \frac{2\sqrt{p-1}ab}{\gamma\sqrt{n}} + \sqrt{\frac{\ln(1/\eta)}{2n}}.$$

Compare to covering number bound: no $\ln n$ but cannot achieve $O(1/n)$ rate.

L_0 **Regularization**

- Only a components of w are nonzeros

$$\hat{w} = \arg\min_w \frac{1}{n} \sum_i I(w^T X_i Y_i \leq 0), \qquad \text{s.t. } \|w\|_0 \leq a.$$

  - more interpretable results
  - good generalization bound in terms of sparsity

Generalization for L_0 regularization

- For each fixed subset of a nonzero coefficients, Sauer's lemma implies infinity-covering of at most $(en/(a+1))^(a+1)$.
- There are only $C_d^a \leq d^a$ possible choices of subset of nonzero coefficients
- In summary, empirical covering is no more than

$$\ln N_\infty(\mathcal{H}, 0|S_n) \leq a \ln d + (a+1)\ln(en/(a+1)).$$

- Implies statistical complexity of $a \ln d/n$
  - applicable even when $d \gg n$:
  - sparsity-level times 1-dimensional complexity (standard for L_0)

General Linear Regularization

- Goal: minimize the average loss $\phi(w^T x), y)$ over unseen data.

- A practical method: minimize observed loss:

$$\hat{w} = \arg\min_w \frac{1}{n} \sum_i \phi(w^T X_i), Y_i), \qquad \text{s.t. } g(w) \leq b.$$

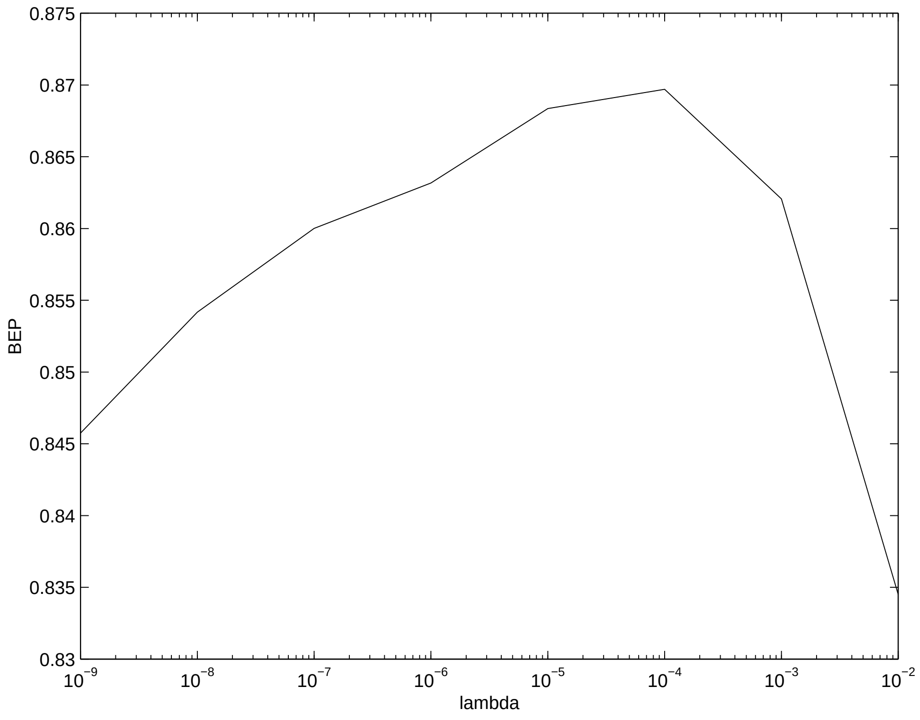
- Equivalent formulation ($\lambda \geq 0$):

$$\hat{w} = \arg\min_w \frac{1}{n} \sum_i \phi(w^T X_i), Y_i) + \lambda g(w).$$

- require convex ϕ and g for computational efficiency.

Effect of Regularization

- Learning complexity controlled by λ: test accuracy versus λ



What Regularization to use

- $\|w\|_2$: when 2-norm of the true classifier is bounded and 2-norm of x is bounded.
- $\|w\|_1$: when 1-norm of the true classifier is bounded and ∞-norm of x is bounded.
  - induce sparse weights (only small number of nonzero weights)
  - automatic feature selection
  - closest convex approximation (relaxation) to L_0 regularization:
- $\|w\|_0$: sparsity with good generalization bound, but non-convex (computionally infeasible).
  - current research: does L_1 relaxaton gives similar generalization performance in terms of sparsity?

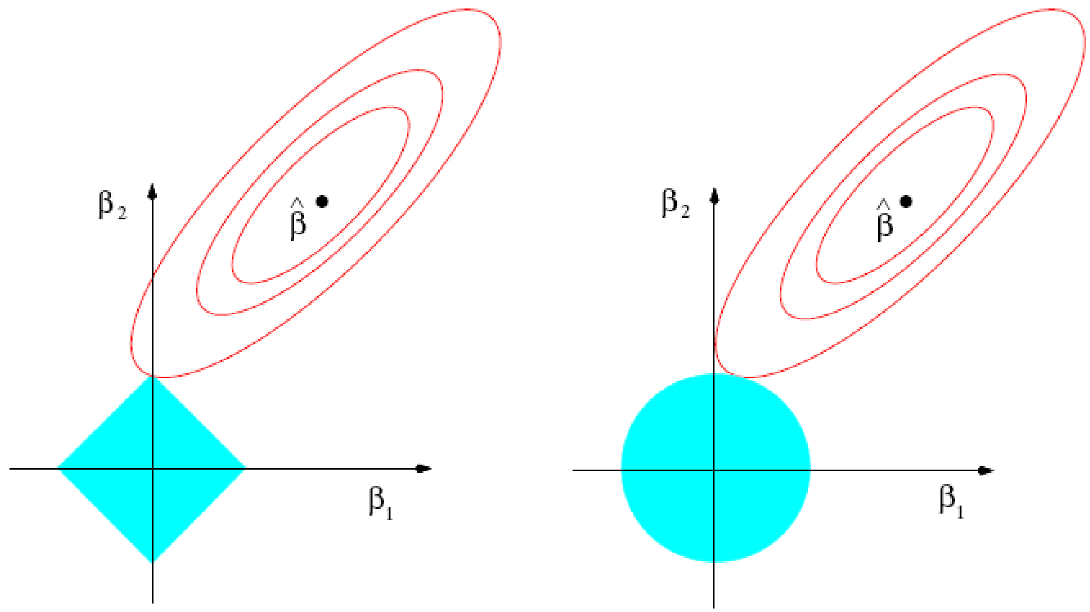


Figure 3.12: *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

Regularization and Stability

- If the loss function is convex, and regularization condition is strictly convex then the regularized solution is stable.
  - adding or removing one component does not change solution much
- Stability leads to good generalization performance: another approach to derive learning bound
  - McDiarmid's inequalit requires stability — stability implies concentration

An example of stability analysis

- Let $w_* = \arg\min_w [E\phi(w^T X, Y) + \lambda w^2]$ be the true parameter
- Let $\hat{w} = \arg\min_w [\frac{1}{n}\sum_i \phi(w^T X_i, Y_i) + \lambda w^2]$ be the estimated estimated parameter.
- Claim (numerical stability): if ϕ is convex in w, then let $M = \sup |\phi_1'(w^T X, Y)| \|X\|_2$, then with probability $1 - \eta$:

$$\|\hat{w} - w_*\|_2 \leq M[1 + \sqrt{2\ln(1/\eta)}]/(\lambda\sqrt{n}).$$

  – this stability result implies good generalization performance:

$$E\phi(\hat{w}^T X, Y) \approx E\phi(w_*^T X, Y).$$

Proof

From

$$\frac{1}{n}\sum_i \phi(\hat{w}^T X_i, Y_i) + \lambda \hat{w}^2 \le \frac{1}{n}\sum_i \phi(w_*^T X_i, Y_i) + \lambda w_*^2,$$

we have

$$\frac{1}{n}\sum_i \underbrace{(\phi(\hat{w}^T X_i, Y_i) - \phi(w_*^T X_i, Y_i) - \phi_1'(w_*^T X_i, Y_i) X_i^T(\hat{w} - w_*))}_{\ge 0}$$

$$+ \lambda \underbrace{(\hat{w}^2 - w_*^2 - 2w_*^T(\hat{w} - w_*))}_{(\hat{w}-w_*)^2}$$

$$\le -(\frac{1}{n}\sum_i \phi_1'(w_*^T X_i, Y_i) X_i + 2\lambda w_*)^T(\hat{w} - w_*)$$

Thus

$$\lambda\|\hat{w}-w_*\|_2^2 \leq \|\frac{1}{n}\sum_i \phi_1'(w_*^T X_i, Y_i)X_i + 2\lambda w_*\|_2 \|\hat{w}-w_*\|_2.$$

Since $E\phi_1'(w_*^T X, Y)X + 2\lambda w_* = 0$, we have

$$\lambda\|\hat{w}-w_*\|_2 \leq \|\frac{1}{n}\sum_i \phi_1'(w_*^T X_i, Y_i)X_i - E\phi_1'(w_*^T X), Y)X\|_2$$

Now apply McDiarmid's inequality, we have with probability $1-\eta$:

$$\lambda\|\hat{w}-w_*\|_2 \leq E\|\frac{1}{n}\sum_i \phi_1'(w_*^T X_i, Y_i)X_i - E\phi_1'(w_*^T X), Y)X\|_2 + M\sqrt{2\ln(1/\eta)/n}$$

$$\leq E^{1/2}\|\frac{1}{n}\sum_i \phi_1'(w_*^T X_i, Y_i)X_i - E\phi_1'(w_*^T X), Y)X\|_2^2 + M\sqrt{2\ln(1/\eta)/n}$$

$$\leq E^{1/2}\sum_i \|\frac{1}{n}\phi_1'(w_*^T X_i, Y_i)X_i\|_2^2 + M\sqrt{2\ln(1/\eta)/n} \leq M(1+\sqrt{2\ln(1/\eta)})/\sqrt{n}.$$

References

- L_∞ covering number bounds for linear regularization:

  T. Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

- Rademacher complexity bounds for linear regularization:

  R. Meir and T. Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.