# High-dimensional Generalized Linear Models and the LASSO

Sara A. Van de Geer

Bin Dai

Department of Statistics,
University of Wisconsin Madison

February 26, 2010

# Outline

# LASSO estimator in generalized linear models

## Linear predictor

Let $Y \in \mathcal{Y} \subset \mathbf{R}$ be a real-valued (response) variable and $X$ be a co-variable with values in some space $\mathcal{X}$. Let

$$\mathcal{F} = \left\{ f_\theta(\cdot) = \sum_{k=1}^{m} \theta_k \psi_k(\cdot),\ \theta \in \Theta \right\}$$

be a (subset of a) linear space of functions on $\mathcal{X}$. Further let $\Theta$ be a convex subset of $\mathbf{R}^m$, possibly $\Theta = \mathbf{R}^m$. The functions $\{\psi_k\}_{k=1}^m$ form a given system of real-valued base functions on $\mathcal{X}$.

# Lasso estimator in generalized linear models

Let $\gamma_f : \mathcal{X} \times \mathcal{Y} \to \mathbf{R}$ be some loss function, and let $\{(X_i, Y_i)\}_{i=1}^n$ be i.i.d. copies of $(X, Y)$. Consider the estimator with lasso penalty

$$\hat{\theta}_n = \text{argmin}_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{u=1}^{n} \gamma_{f_\theta}(X_i, Y_i) + \lambda_n \hat{I}(\theta) \right\},$$

where

$$\hat{I}(\theta) := \sum_{k=1}^{m} \hat{\sigma}_k |\theta_k|$$

denotes the weighted $l_1$ norm of the vector $\theta \in \mathbf{R}^m$, with random weights

$$\hat{\sigma}_k := \left( \frac{1}{n} \sum_{i=1}^{n} \psi_k^2(X_i) \right)^{1/2}$$

# Goal of this paper

## The best linear predictor

Let $P$ be the distribution of $(X, Y)$. The target function $\bar{f}$ is defined as

$$\bar{f} := argmin_{f \in \mathbf{F}} P_{\gamma_f},$$

where $F \supseteq \mathcal{F}$ (and assuming for simplicity that there is a unique minimum). It will be shown that if the target $\bar{f}$ can be well approximated by a sparse function $f_{\theta_n^*}$, the estimator $\hat{\theta}_n$ will have prediction error roughly as if it knew this sparseness.

The excess risk of $f$ is

$$\mathcal{E}(f) := P_{\gamma_f} - P_{\gamma_{\bar{f}}}$$

A probability inequality will be derived for the excess risk $\mathcal{E}(f_{\hat{\theta}_n})$.

# Outline

# Assumptions

## Assumption L

The loss function $\gamma_f$ is of the form $\gamma_f(x, y) = \gamma(f(x), y) + b(f)$, where $b(f)$ is a constant which is convex in $f$, and $\gamma(\cdot, y)$ is convex for all $y \in \mathcal{Y}$. Moreover, it satisfies the Lipschitz property

$$\begin{aligned} |\gamma(f_\theta(x), y) - \gamma(f_{\bar\theta}(x), y)| \quad &\le \quad |f_\theta(x) - f_{\bar\theta}(x)| \\ &\qquad \forall (x, y) \in \mathcal{X} \times \mathcal{Y},\ \forall \theta, \bar\theta \in \Theta. \end{aligned}$$

## Assumption A

It holds that

$$K_m := \max_{1 \le k \le m} \frac{||\psi_k||_\infty}{\sigma_k} < \infty$$

# Assumptions

## Assumption B

There exists an $\eta > 0$ and strictly convex increasing $G$, such that for all $\theta \in \Theta$ with $||f_\theta - \bar{f}||_\infty \leq \eta$, one has

$$\mathcal{E}(f_\theta) \geq G(||f_\theta - \bar{f}||).$$

## Assumption C

There exists a function $D(\cdot)$ on the subsets of the index set $\{1, \ldots, m\}$, such that for all $\mathcal{K} \subset \{1, \ldots, m\}$, and for all $\theta \in \Theta$ and $\tilde{\theta} \in \Theta$, we have

$$\sum_{k \in \mathcal{K}} \sigma_k |\theta_k - \tilde{\theta}_k| \leq \sqrt{D(\mathcal{K})} ||f_\theta - f_{\tilde{\theta}}||.$$

$$D_\theta := D(\{k : |\theta_k| \neq 0\}).$$

# Further quantities

The convex conjugate of the function $G$ given in Assumption B is denoted $H$.

## Smoothing parameter

Let

$$\bar{a}_n = 4a_n, \quad a_n := \left(\sqrt{\frac{2\log(2m)}{n}} + \frac{\log(2m)}{n} K_m\right)$$

Further let for $t > 0$,

$$\lambda_{n,0} := \lambda_{n,0}(t) \quad := \quad a_n \left(1 + t\sqrt{2(1 + 2a_n K_m)} + \frac{2t^2 a_n K_m}{3}\right)$$

$$\bar{\lambda}_{n,0} := \bar{\lambda}_{n,0}(t) \quad := \quad \bar{a}_n \left(1 + t\sqrt{2(1 + 2\bar{a}_n K_m)} + \frac{2t^2 \bar{a}_n K_m}{3}\right)$$

# Penalty Function

Let

$$I(\theta) := \sum_{k=1}^{m} \sigma_k |\theta_k|.$$

and $\hat{I}(\theta) = \sum_{k=1}^{m} \hat{\sigma}_k |\theta_k|$ its empirical $l_1$ norm. Moreover, for any $\theta$ and $\tilde{\theta}$ in $\Theta$, let

$$I_1(\theta|\tilde{\theta}) := \sum_{k:\tilde{\theta}_k \neq 0} \sigma_k |\theta_k|, \quad I_2(\theta|\tilde{\theta}) := I(\theta) - I_1(\theta|\tilde{\theta}).$$

Likewise for the empirical versions:

$$\hat{I}_1(\theta|\tilde{\theta}) := \sum_{k:\tilde{\theta}_k \neq 0} \hat{\sigma}_k |\theta_k|, \quad \hat{I}_2(\theta|\tilde{\theta}) := \hat{I}(\theta) - \hat{I}_1(\theta|\tilde{\theta}).$$

# Outline

# Nonrandom Normalization Weights in the Penalty

## Quantities

1. $\lambda_n := 2\bar{\lambda}_{n,0}$,
2. $\mathcal{V}_\theta := H(4\lambda_n\sqrt{D_\theta})$ (estimation error),
3. $\theta_n^* := arg\ \min_{\theta\in\Theta}\{\mathcal{E}(f_\theta) + \mathcal{V}_\theta\}$ (oracle),
4. $2\epsilon_n^* := 3\mathcal{E}(f_{\theta_n^*}) + 2\mathcal{V}_{\theta_n^*}$ (oracle rate),
5. $\zeta_n^* := \epsilon_n^*/\bar{\lambda}_{n,0}$ (oracle rate for $l_1$),
6. $\theta(\epsilon_n^*) := arg\ \min_{\theta\in\Theta, I(\theta-\theta_n^*)\leq 6\zeta_n^*}\{\mathcal{E}(f_\theta) - 4\lambda_n I_1(\theta - \theta_n^*|\theta_n^*)\}$.

## Conditions

1. It holds that $||f_{\theta_n^*} - \bar{f}||_\infty \leq \eta$, where $\eta$ is given in Assumption B.
2. It holds that $||f_{\theta(\epsilon_n^*)} - \bar{f}||_\infty \leq \eta$, where $\eta$ is given in Assumption B.

# Nonrandom Normalization Weights in the Penalty

## THEOREM 2.1

Suppose Assumptions L, A, B and C, and Conditions I and II hold. Let $\lambda_n$, $\theta_n^*$, $\epsilon_n^*$ and $\zeta_n^*$ be given. Assume $\sigma_k$ is known for all $k$ and let $\hat{\theta}_n$ be the lasso estimator. Then we have with probability at least

$$1 - 7\exp[-n\bar{a}_n^2 t^2],$$

that

$$\mathcal{E}(f_{\hat{\theta}_n}) \leq 2\epsilon_n^*,$$

and moreover

$$2I(\hat{\theta}_n - \theta_n^*) \leq 7\zeta_n^*.$$

# Random Normalization Weights in the Penalty

## Quantities

1. $\lambda_n := 3\bar{\lambda}_{n,0}$,
2. $\mathcal{V}_\theta := H(5\lambda_n\sqrt{D_\theta})$,
3. $\theta_n^* := arg\,\min_{\theta\in\Theta}\{\mathcal{E}(f_\theta) + \mathcal{V}_\theta\}$,
4. $2\epsilon_n^* := 3\mathcal{E}(f_{\theta_n^*}) + 2\mathcal{V}_{\theta_n^*}$,
5. $\zeta_n^* := \epsilon_n^*/\bar{\lambda}_{n,0}$,
6. $\theta(\epsilon_n^*) := arg\,\min_{\theta\in\Theta, I(\theta-\theta_n^*)\le 6\zeta_n^*}\{\mathcal{E}(f_\theta) - 5\lambda_n I_1(\theta - \theta_n^*|\theta_n^*)\}$.

## Conditions

1. Conditions I and II in nonrandom normalization case.
2. $\sqrt{\frac{\log(2m)}{n}}K_m \le 0.13$.

# Nonrandom Normalization Weights in the Penalty

## THEOREM 2.2

Suppose Assumptions L, A, B and C, and Conditions I, II and III hold. Let $\lambda_n$, $\theta_n^*$, $\epsilon_n^*$ and $\zeta_n^*$ be given, and the weights $\hat{\sigma}_k$ should be estimated. Take $\bar{\lambda}_{n,0} > 4\sqrt{\frac{\log(2m)}{n}} \times (1.6)$ Then with probability at least $1-\alpha$, we have that

$$\mathcal{E}(f_{\hat{\theta}_n}) \leq 2\epsilon_n^*,$$

and moreover

$$2I(\hat{\theta}_n - \theta_n^*) \leq 7\zeta_n^*.$$

Here $\alpha = \exp[-na_n^2 s^2] + 7\exp[-n\bar{a}_n^2 t^2]$, with $s > 0$ being defined by $\frac{5}{9} = K_m \lambda_{n,0}(s)$, and $t > 0$ being defined by $\bar{\lambda}_{n,0} = \bar{\lambda}_{n,0}(t)$.

# Outline

# Loss functions

## Example of loss functions satisfying Assumptions L, B

- Logistic Regression

$$\gamma_f(x,y) = [-f(x)y + \log(1+\exp(f(x))]/2$$

- Density estimation
- Hinge loss for support vector machine

$$\gamma_f(x,y) = (1 - yf(x))_+.$$

However, the usual quadratic loss is not Lipschitz on the whole real line.

# Theorem 3.1

Suppose Assumptions A and C hold. Let $\lambda_n$, $\theta_n^*$, $\epsilon_n^*$ and $\zeta_n^*$ be given, with $H(v) = v^2/2$, $v > 0$, but now with $\bar{\lambda}_{n,0}$ replaced by

$$\tilde{\lambda}_{n,0} := \sqrt{\frac{14}{9}}\sqrt{\frac{2\log(2m)}{n} + 2t^2\bar{a}_n^2} + \bar{\lambda}_{n,0}.$$

Assume moreover that $||f_{\theta_n^*} - \bar{f}||_\infty \leq \eta \leq 1/2$, that $6\zeta_n^* K_m + 2\eta \leq 1$, and that $\sqrt{\frac{\log(2m)}{n}} K_m \leq 0.33$. Let $\sigma_k$ be known for all $k$ and let $\hat{\theta}_n$ be the lasso estimator. Then with probability at least $1 - \alpha$, that

$$\begin{aligned} \mathcal{E}(f_{\hat{\theta}_n}) &\leq 2\epsilon_n^* \\ 2I(\hat{\theta}_n - \theta_n^*) &\leq 7\zeta_n^* \end{aligned}$$

Here $\alpha = \exp[-na_n^2 s^2] + 7\exp[-n\bar{a}_n^2 t^2]$, with $s > 0$ a solution of $\frac{9}{5} = K_m \lambda_{n,0}(s)$.

# Outline

# Concentration theorem

Let $Z_1, \ldots, Z_n$ be independent random variables with values in space $\mathcal{Z}$ and let $\Gamma$ be a class of real-valued functions on $\mathcal{Z}$, satisfying for some positive constants $\eta_n$ and $\tau_n$

$$\begin{aligned} ||\gamma_n||_\infty &\leq \eta_n \;\; \forall \gamma \in \Gamma \\ \frac{1}{n}\sum_{i=1}^{n} var(\gamma(Z_i)) &\leq \tau_n^2 \;\; \forall \gamma \in \Gamma. \end{aligned}$$

Define

$$\mathbf{Z} := \sup_{\gamma \in \Gamma} |\frac{1}{n}\sum_{i=1}^{n}(\gamma(Z_i) - E\gamma(Z_i))|.$$

Then for $z > 0$,

$$\mathbf{P}\left(\mathbf{Z} \geq E\mathbf{Z} + z\sqrt{2(\tau_n^2 + 2\eta_n E\mathbf{Z})} + \frac{2z^2\eta_n}{3}\right) \leq \exp[-nz^2].$$

# Symmetrization theorem

## Rademacher sequence

i.i.d. random variables $\epsilon_1, \ldots, \epsilon_n$, taking values $\pm 1$ each with probability $1/2$.

Let $Z_1, \ldots, Z_n$ be independent random variables with values in $\mathcal{Z}$, and let $\epsilon_1, \ldots, \epsilon_n$ be a Rademacher sequence independent of $Z_1, \ldots, Z_n$. Let $\Gamma$ be a class of real-valued functions on $\mathcal{Z}$. Then

$$E\left(\sup_{\gamma\in\Gamma}\left|\sum_{i=1}^{n}\{\gamma(Z_i)-E\gamma(Z_i)\}\right|\right)\leq 2E\left(\sup_{\gamma\in\Gamma}\left|\sum_{i=1}^{n}\epsilon_i\gamma(Z_i)\right|\right).$$

# Contraction theorem

Let $z_1, \ldots, z_n$ be nonrandom elements of some space $\mathcal{Z}$ and let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{Z}$. Consider Lipschitz function $\gamma_i : \mathbf{R} \to \mathbf{R}$, that is,

$$|\gamma_i(s) - \gamma_i(\tilde{s})| \leq |s - \tilde{s}| \quad \forall\, s, \tilde{s} \in \mathbf{R}$$

Let $\epsilon_1, \ldots, \epsilon_n$ be a Rademacher sequence. Then for any function $f^* : \mathcal{Z} \to \mathbf{R}$, we have

$$E\left(\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \epsilon_i \{\gamma_i(f(z_i)) - \gamma_i(f^*(z_i))\} \right| \right) \leq 2E\left(\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \epsilon_i (f(z_i) - f^*(z_i)) \right| \right).$$

# Lemma A.1

Let $Z_1, \ldots, Z_n$ be independent $\mathcal{Z}$-valued random variables, and $\gamma_1, \ldots, \gamma_n$, be real-valued functions on $\mathcal{Z}$, satisfying for $k = 1, \ldots, m$,

$$E\gamma_k(Z_i) = 0, \ \forall \ i \quad ||\gamma_k||_\infty \leq \eta_n, \ \frac{1}{n}\sum_{i=1}^{n} E\gamma_k^2(Z_i) \leq \tau_n^2.$$

Then

$$E\left(\max_{1\leq k\leq m}\left|\frac{1}{n}\sum_{i=1}^{n}\gamma_k(Z_i)\right|\right) \leq \sqrt{\frac{2\tau_n^2\log(2m)}{n}} + \frac{\eta_n\log(2m)}{n}.$$

# Outline

# Lemma A.2

Let $\epsilon_1, \dots, \epsilon_n$ be Rademacher sequence, independent of the training set $(X_1, Y_1), \dots, (X_n, Y_n)$. Moreover, fix some $\theta^* \in \Theta$ and let for $M > 0$, $\mathcal{F}_M := \{f_\theta : \theta \in \Theta, I(\theta - \theta^*) \leq M\}$ and

$$\mathbf{Z}(M) := \sup_{f \in \mathcal{F}_M} |(P_n - P)(\gamma_{f_\theta} - \gamma_{f_{\theta^*}})|,$$

We have

$$E\mathbf{Z}(M) \leq 4ME \left( \max_{1 \leq k \leq m} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \psi_k(X_i)/\sigma_k \right| \right)$$

# Proof of Lemma A.2

$$E\mathbf{Z}(M) \le 2E\left(\sup_{f\in\mathcal{F}_M}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\{\gamma(f_\theta(X_i),Y_i)-\gamma(f_{\theta^*}(X_i),Y_i)\}\right|\right)$$

$$E_{(X,Y)}\left(\sum_{f\in\mathcal{F}_M}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\{\gamma(f_\theta(X_i),Y_i)-\gamma(f_{\theta^*}(X_i),Y_i)\}\right|\right)$$

$$\le 2E_{(X,Y)}\left(\sup_{f\in\mathcal{F}_M}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(f_\theta(X_i)-f_{\theta^*}(X_i))\right|\right)$$

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(f_\theta(X_i)-f_{\theta^*}(X_i))\right| &\le \sum_{k=1}^{m}\sigma_k|\theta_k-\theta^*|\max_{1\le k\le m}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\psi_k(X_i)/\sigma\right. \\ &= I(\theta-\theta^*)\max_{1\le k\le m}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\psi_k(X_i)/\sigma_k\right|.\end{aligned}$$

# Lemma A.3

The distribution of $X$ is denoted by $Q$, and the empirical distribution of covariates $\{X_i\}_{i=1}^n$ is written as $Q_n$.

## Statement

We have

$$E\left(\max_{1\le k\le m}\left|\frac{(Q_n-Q)(\psi_k)}{\sigma_k}\right|\right) \le a_n,$$
$$E\left(\max_{1\le k\le m}\frac{|1/n\sum_{i=1}^n \epsilon_i\psi(X_i)|}{\sigma_k}\right) \le a_n.$$

**Proof**: This follows from $||\psi_k||_\infty/\sigma_k \le K_m$ and $var(\psi_k(X))/\sigma_k^2 \le 1$. So apply Lemma A.1 with $\eta_n = K_m$ and $\tau_n^2 = 1$.

# Corollary A.1

For all $M > 0$ and all $\theta \in \Theta$ with $I(\theta - \theta^*) \leq M$, it holds that

$$||\gamma_{f_\theta} - \gamma_{f_{\theta*}}||_\infty \leq MK_m$$
$$P(\gamma_{f_\theta} - \gamma_{f_{\theta*}})^2 \leq M^2.$$

Therefore, since by Lemma A.2 and Lemma A.3, for all $M > 0$,

$$\frac{E\mathbf{Z}(M)}{M} \leq \bar{a}_n, \quad \bar{a}_n = 4a_n,$$

we have, in view of Bousquet's Concentration theorem, for all $M > 0$ and all $t > 0$,

$$\mathbf{P}\left(\mathbf{Z}(M) \geq \bar{a}_n M\left(1 + t\sqrt{2(1 + 2\bar{a}_n K_m)} + \frac{2t^2\bar{a}_n K_m}{3}\right)\right) \leq \exp[-n\bar{a}_n^2 t^2].$$

# Outline

## A general theorem of nonrandom weights

Take $b > 0, d > 1$, and $d_b := d\left(\frac{b+d}{(d-1)b} \vee 1\right)$.

**Quantities**:

1. $\lambda_n := (1+b)\bar{\lambda}_{n,0}$,
2. $\mathcal{V}_\theta := 2\delta H(\frac{2\lambda_n\sqrt{D_\theta}}{\delta})$, where $0 < \delta < 1$,
3. $\theta_n^* := arg\min_{\theta\in\Theta}\{\mathcal{E}(f_\theta) + \mathcal{V}_\theta\}$ ,
4. $\epsilon_n^* := (1+\delta)\mathcal{E}(f_{\theta_n^*}) + \mathcal{V}_{\theta_n^*}$,
5. $\zeta_n^* := \frac{\epsilon_n^*}{\bar{\lambda}_{n,0}}$,
6. $\theta(\epsilon_n^*) := arg\min_{\theta\in\Theta, I(\theta-\theta_n^*)\le d_b\zeta_n^*/b}\{\delta\mathcal{E}(f_\theta) - 2\lambda_n I_1(\theta-\theta_n^*|\theta_n^*)\}$.

Conditions same as in Theorem 2.1. Theorem 2.1 is the special case with $b = 1$, $\delta = 1/2$ and $d = 2$.

# Lemma A.4

## Statement

Suppose conditions are met. For all $\theta \in \Theta$ with $I(\theta - \theta_n^*) \leq d_b \zeta_n^* / b$, it holds that

$$2\lambda_n I_1(\theta - \theta_n^* | \theta_n^*) \leq \delta \mathcal{E}(f_\theta) + \epsilon_n^* - \mathcal{E}(f_{\theta_n^*}).$$

**Proof**:

$$\begin{aligned} 2\lambda_n I_1(\theta - \theta_n^*) &= 2\lambda_n I_1(\theta - \theta_n^*) - \delta \mathcal{E}(f_\theta) + \delta \mathcal{E}(f_\theta) \\ &\leq 2\lambda_n I_1(\theta(\epsilon_n^*) - \theta_n^*) - \delta \mathcal{E}(f_{\theta(\epsilon^*)}) + \delta E(f_\theta). \end{aligned}$$

By Assumption C, and Condition II,

$$2\lambda_n I_1(\theta(\epsilon_n^*) - \theta_n^*) \leq 2\lambda_n \sqrt{D_{\theta_n^*}} || f_{\theta(\epsilon_n^*)} - f_{\theta_n^*} ||.$$

# Proof of Lemma A.4 (cont'd)

By the triangle inequality,

$$2\lambda_n\sqrt{D_{\theta_n^*}}||f_{\theta(\epsilon_n^*)} - f_{\theta_n^*}|| \leq 2\lambda_n\sqrt{D_{\theta_n^*}}||f_{\theta(\epsilon_n^*)} - \bar{f}|| + 2\lambda_n\sqrt{D_{\theta_n^*}}||f_{\theta(\epsilon_n^*)} - \bar{f}||.$$

It follows from conditions I and II, combined with Assumption B, that

$$2\lambda_n\sqrt{D_{\theta_n^*}}||f_{\theta(\epsilon_n^*)} - f_{\theta_n^*}|| \leq \delta\mathcal{E}(f_{\theta(\epsilon_n^*)}) + \delta\mathcal{E}(f_{\theta_n^*}) + \mathcal{V}_{\theta_n^*}.$$

Hence, when $I(\theta - \theta_n^*) \leq d_b\zeta_n^*/b$,

$$\begin{aligned} 2\lambda_n I_1(\theta - \theta_n^*) &\leq \delta\mathcal{E}(f_\theta) + \delta\mathcal{E}(f_{\theta_n^*}) + \mathcal{V}_{\theta_n^*} \\ &= \delta\mathcal{E}(f_\theta) + \epsilon_n^* - \mathcal{E}(f_{\theta_n^*}). \end{aligned}$$

# Lemma A.5

Suppose Conditions I and II are met. Consider any (random) $\tilde{\theta} \in \Theta$ with $R_n(f_{\tilde{\theta}}) + \lambda_n I(\tilde{\theta}) \leq R_n(f_{\theta_n^*}) + \lambda_n I(\theta_n^*)$. Let $1 < d_0 \leq d_b$. Then

$$\mathbf{P}\left( I(\tilde{\theta} - \theta_n^*) \leq d_n \frac{\zeta_n^*}{b} \right) \leq \mathbf{P}\left( I(\tilde{\theta} - \theta_n^*) \leq \left( \frac{d_0 + b}{1 + b} \right) \frac{\zeta_n^*}{b} \right) + \exp[-n\bar{a}_n^2 t^2]$$

**Proof**: Let $\tilde{\mathcal{E}} := \mathcal{E}(f_{\tilde{\theta}})$ and $\mathcal{E}^* := \mathcal{E}(f_{\theta_n^*})$. Since $R_n(f_{\tilde{\theta}}) + \lambda_n I(\tilde{\theta}) \leq R_n(f_{\theta_n^*}) + \lambda_n I(\theta_n^*)$, and known $I(\tilde{\theta} - \theta_n^*) \leq d_0 \zeta_n^*/b$, that

$$\tilde{\mathcal{E}} + \lambda_n I(\tilde{\theta}) \leq \mathbf{Z}(d_0 \zeta_n^*/b) + \mathcal{E}^* + \lambda_n I(\theta_n^*).$$

With probability at least $1 - \exp[-n\bar{a}_n^2 t^2]$, the random variable $\mathbf{Z}(d_0\zeta_n^*/b)$ is bounded by $\bar{\lambda}_{n,0} d_0 \zeta_n^*/b$. But we then have

$$\tilde{\mathcal{E}} + \lambda_n I(\tilde{\theta}) \leq \bar{\lambda}_{n,0} d_0 \zeta_n^*/b + \mathcal{E}^* + \lambda_n I(\theta_n^*).$$

# Proof of Lemma A.5 (cont'd)

Then on event $\{I(\tilde{\theta}-\theta_n^*) \le d_0\zeta_n^*/b\} \cup \{\mathbf{Z}(d_0\zeta_n^*/b) \le \bar{\lambda}_{n,0} d_0\zeta_n^*/b\}$, invoking $\lambda_n = (1+b)\bar{\lambda}_{n,0}$, $I(\tilde{\theta}) = I_1(\tilde{\theta}) + I_2(\tilde{\theta})$ and $I(\theta_n^*) = I_1(\theta_n^*)$, that

$$\tilde{\mathcal{E}} + (1+b)\bar{\lambda}_{n,0} I_2(\tilde{\theta}) \le \bar{\lambda}_{n,0}\frac{d_0\zeta_n^*}{b} + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0} I_1(\tilde{\theta}-\theta_n^*).$$

But $I_2(\tilde{\theta}) = I_2(\tilde{\theta}-\theta_n^*)$. So if add another $(1+b)\bar{\lambda}_{n,0} I_1(\tilde{\theta}-\theta_n^*)$ to both sides of the last inequality, we obtain

$$\begin{aligned}
\tilde{\mathcal{E}} + (1+b)\bar{\lambda}_{n,0} I(\tilde{\theta}-\theta_n^*) &\le \bar{\lambda}_{n,0}\frac{d_0\zeta_n^*}{b} + 2(1+b)\bar{\lambda}_{n,0} I_1(\tilde{\theta}-\theta_n^*) + \mathcal{E}^* \\
&\le \bar{\lambda}_{n,0}\frac{d_0\zeta_n^*}{b} + \delta\tilde{\mathcal{E}} + \epsilon_n^* \\
&= (d_0+b)\bar{\lambda}_{n,0}\frac{\zeta_n^*}{b} + \delta\tilde{\mathcal{E}},
\end{aligned}$$

The result follows as $\epsilon_n^* = \bar{\lambda}_{n,0}\zeta_n^*$ and $0 < \delta < 1$.

# Corollary A.2 and Lemma A.6

**Corollary A.2**: Suppose conditions I and II are met. Let $d_0 \le d_b$. For any (random) $\tilde{\theta} \in \Theta$ with $R_n(f_{\tilde{\theta}}) + \lambda_n I(\tilde{\theta}) \le R_n(f_{\theta_n^*}) + \lambda_n I(\theta_n^*)$,

$$\mathbf{P}\left( I(\tilde{\theta} - \theta_n^*) \le d_n \frac{\zeta_n^*}{b} \right)$$

$$\le \mathbf{P}\left( I(\theta - \theta) \le (1 + (d_0 + 1)(1 + b)^{-N}) \frac{\zeta_n^*}{b} \right) + \exp[-n\bar{a}_n^2 t^2].$$

**Lemma A.6**: Suppose conditions I and II are met, define

$$\tilde{\theta}_s = s\hat{\theta}_n + (1 - s)\theta_n^*$$

$$s = \frac{d\zeta_n^*}{d\zeta_n^* + bI(\hat{\theta}_n - \theta_n^*}.$$

Then for any integer N, with probability $1 - N\exp[-n\bar{a}_n^2 t^2]$ we have

$$I(\tilde{\theta}_s - \theta_n^*) \le \left(1 + (d - 1)(1 + b)^{-N}\right) \frac{\zeta_n^*}{b}.$$

# Lemma A.7

## Statement

Suppose conditions I and II are met. Let $N_1 \in \mathbf{N}$ and $N_2 \in \mathbf{N} \cup \{0\}$. Define $\delta_1 = (1+b)^{-N_1}$ $(N_1 \geq 1)$, and $\delta_2 = (1+b)^{-N_2}$. With probability at least $1 - (N_1 + N_2)\exp[-n\bar{a}_n^2 t^2]$, we have

$$I(\hat{\theta}_n - \theta_n^*) \leq d(\delta_1, \delta_2)\frac{\zeta_n^*}{b},$$

with

$$d(\delta_1, \delta_2) = 1 + \left(\frac{1 + (d^2 - 1)\delta_1}{(d-1)(1-\delta_1)}\right)\delta_2.$$

# Outline

# Theorem A.4

Write

$$\triangle(b,\delta,\delta_1,\delta_2) := d(\delta_1,\delta_2)\frac{1-\delta^2}{\delta b} \vee 1.$$

Suppose condition I and II are met. Let $\delta_1$ and $\delta_2$ as in Lemma A.7. We have the probability at least

$$1-\left(\log_{1+b}\frac{(1+b)^2 \triangle (b,\delta,\delta_1,\delta_2)}{\delta_1\delta_2}\right)\exp[-n\bar{a}_n^2 t^2],$$

that

$$\mathcal{E}(f_{\hat{\theta}_n}) \leq \frac{\epsilon_n^*}{1-\delta},$$

$$I(\hat{\theta}_n-\theta_n^*) \leq d(\delta_1,\delta_2)\frac{\zeta_n^*}{b}.$$

# Proof of theorem A.4

Define $\hat{\mathcal{E}} := \mathcal{E}(\hat{f}_{\hat{\theta}_n})$ and $\mathcal{E}^* := \mathcal{E}(f_{\theta_n^*})$. Set $c := \frac{\delta b}{1-\delta^2}$, we consider the cases (a) $c < d(\delta_1, \delta_2)$ and (b) $c \geq d(\delta_1, \delta_2)$.

**(a)**: Suppose that first $c < d(\delta_1, \delta_2)$. Let $J$ be an integer satisfying $(1+b)^{J-1} c \leq d(\delta_1, \delta_2)$ and $(1+b)^J c > d(\delta_1, \delta_2)$.

Consider two cases:

**(a1)** If $c\zeta_n^*/b < I(\hat{\theta}_n - \theta_n^*) \leq d(\delta_1, \delta_2)\zeta_n^*/b$, then

$$(1+b)^{j-1} c\zeta_n^*/b < I(\hat{\theta}_n - \theta_n^*) \leq (1+b)^j c\zeta_n^*/b$$

for some $j \in \{1, \dots, J\}$. Expect on set with probability at most $\exp[-n\bar{a}_n^2 t^2]$, we thus have

$$\hat{\mathcal{E}} + (1+b)\bar{\lambda}_{n,0} I(\hat{\theta}_n) \leq (1+b)\bar{\lambda}_{n,0} I(\hat{\theta}_n - \theta_n^*) + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0} I(\theta_n^*).$$

So then by similar arguments as in the proof of Lemma A.5,

$$\hat{\mathcal{E}} \leq 2(1+b)\bar{\lambda}_{n,0} I_1(\hat{\theta}_n - \theta_n^*) + \mathcal{E}^*.$$

Since $d(\delta_1, \delta_2) \leq d_b$, we obtain $\hat{\mathcal{E}} \leq \epsilon_n^* + \delta\hat{\mathcal{E}}$ so then $\hat{\mathcal{E}} \leq \frac{\epsilon_n^*}{1-\delta}$.

# Proof of theorem A.4 (cont'd)

**(a2)** If $I(\hat{\theta}_n - \theta_n^*) \leq c\zeta_n^*/b$, except on a set with probability at most $\exp[-n\bar{a}_n^2 t^2]$, that

$$\hat{\mathcal{E}} + (1+b)\bar{\lambda}_{n,0} I(\hat{\theta}_n) \leq \left(\frac{\delta}{1-\delta^2}\right)\bar{\lambda}_{n,0}\zeta_n^* + \mathcal{E}^* + (1+b)I(\theta_n^*), \quad (1)$$

Which gives

$$\begin{aligned}
\hat{\mathcal{E}} &\leq \left(\frac{\delta}{1-\delta^2}\right)\bar{\lambda}_{n,0}\zeta_n^* + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0} I_1(\hat{\theta}_n - \theta_n^*) \\
&\leq \left(\frac{\delta}{1-\delta^2}\right)\bar{\lambda}_{n,0}\zeta_n^* + \mathcal{E}^* + \frac{\delta}{2}\mathcal{E}^* + \frac{\mathcal{V}_{\theta_n^*}}{2} + \frac{\delta}{2}\hat{\mathcal{E}} \\
&\leq \left(\frac{\delta}{1-\delta^2} + \frac{1}{2}\right)\epsilon_n^* + \frac{\mathcal{E}^*}{2} + \frac{\delta}{2}\hat{\mathcal{E}}.
\end{aligned}$$

# Proof of theorem A.4 (cont'd)

This yields

$$\hat{\mathcal{E}} \leq \frac{2}{2-\delta}\left(\frac{\delta}{1-\delta^2}+\frac{1}{2}+\frac{1}{2(1+\delta)}\right)\epsilon_n^* = \frac{1}{1-\delta}\epsilon_n^*.$$

Furthermore, by lemma A.7, with probability at least $1-(N_1+N_2)\exp[-n\bar{a}_n^2 t^2]$, that

$$I(\hat{\theta}_n-\theta_n^*) \leq \frac{d(\delta_1,\delta_2)}{b}\zeta_n^*$$

The result follows from

$$J+1 \leq \log_{1+b}\left(\frac{(1+b)^2 d(\delta_1,\delta_2)}{c}\right)$$

$$N_1 = \log_{1+b}\left(\frac{1}{\delta_1}\right) \quad N_2 = \log_{1+b}\left(\frac{1}{\delta_2}\right)$$

**(b)** Finally, consider the case $c \geq d(\delta_1, \delta_2)$.Then on the set where $I(\hat{\theta}_n - \theta_n^*) \leq d(\delta_1, \delta_2)\zeta_n^*/b$, again have that except on a subset with probability at most $\exp[-n\bar{a}_n^2 t^2]$,

$$\begin{aligned} \hat{\mathcal{E}} + (1+b)\bar{\lambda}_{n,0} I(\hat{\theta}_n) &\leq d(\delta_1, \delta_2)\frac{\zeta_n^*}{b} + \mathcal{E}^* + (1+b)I(\theta_n^*) \\ &\leq \left(\frac{\delta}{1-\delta^2}\right)\bar{\lambda}_{n,0}\zeta_n^* + \mathcal{E}^* + (1+b)I(\theta_n^*), \end{aligned}$$

as

$$d(\delta_1, \delta_2) \leq c = \frac{\delta b}{1-\delta_2}.$$

We arrive at the same inequality in (1) and may proceed as there. Note finally that also in this case

$$\begin{aligned} (N_1 + N_2 + 1) &\leq \log_{1+b} \frac{(1+b)^2}{\delta_1 \delta_2} \\ &\leq \log_{1+b} \frac{(1+b)^2 \bigtriangleup (b, \delta, \delta_1, \delta_2)}{\delta_1 \delta_2}. \end{aligned}$$