Impossible of successful classification when useful features are rare and weak

Jiashun Jin

Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213

Communicated by David L. Donoho, Stanford University, Stanford, CA, April 9, 2009 (received for review February 21, 2009)

We study a two-class classification problem with a large number of features, out of which many are useless and only a few are useful, but we do not know which ones they are. The number of features is large compared with the number of training observations. Calibrating the model with 4 key parameters—the number of features, the size of the training sample, the fraction, and strength of usefulness—we identify a region in parameter space where no classifier can reliably separate the two classes on fresh data. The complement of this region—where successful classification is possible—is also briefly discussed.

The subset of useful features, no classifier built from the dataset is likely to separate the two classes on fresh data.

In companion work (5), we show that in the complement of this region, a specific method for classifier training—Higher Criticism Threshold feature selection (4)—does work, and so the results here are definitive.

Classification When Features Are Rare and Weak

Consider a two-class classification setting where we have a set of labeled training samples \((Y_i, X_i), i = 1, 2, \ldots, n\). Each label \(Y_i = 1\) if \(X_i\) comes from class 1 and \(Y_i = -1\) if \(X_i\) comes from class 2, and each feature vector \(X_i \in \mathbb{R}^p\). For simplicity, we suppose that the training set contains equal numbers of samples from each of the two classes, and that the feature vector obeys \(X_i \sim N(\mu, I_p)\), \(i = 1, 2, \ldots, n\), for an unknown mean contrast vector \(\mu \in \mathbb{R}^p\). Also, we suppose that the feature covariance matrix is the identity matrix. Extension to correlated cases is possible if side information about the feature covariance is available (see ref. 6, for example).

Following the two companion papers (4, 5), we consider the following rare/weak feature model (RW model), where the vector \(\mu\) is nonzero in only an \(\epsilon\) fraction of coordinates, and the nonzero coordinates of \(\mu\) share a common amplitude \(\mu_0\). Formally speaking, let \(I_1, I_2, \ldots, I_p\) be samples from Bernoulli(\(\epsilon\)), and let

\[
\mu(j) = \mu_0 \cdot I_j, \quad 1 \leq j \leq p.
\]

Let \(Z\) denote the vector of \(Z\) scores corresponding to the training set: \(Z(j) = (1/\sqrt{n}) \sum_{i=1}^n Y_i \cdot X_{ij}\). The \(j\) th \(Z\) score arises in a formal normal-theory test of whether the \(j\) th feature is useless or useful. Under our assumptions, \(Z \sim N(\mu_0, I_p)\), thus each coordinate of \(Z\) has expectation either 0 or \(\tau = \sqrt{\mu_0}\).

We assume \(p \gg n, \epsilon \) is small, and \(\tau\) is either small or moderately large (e.g., \(p = 10,000, n = 100, \epsilon = 0.01, \tau = 2\)). Because zero coordinates of \(\mu\) are entirely noninformative for classification, the useful features are those with nonzero coordinates in \(\mu\). The parameters \(\epsilon\) and \(\tau\) can be set to make such useful features arbitrarily rare (by setting \(\epsilon\) close to 0) and weak (setting \(\tau\) small); we denote an instance of the rare/weak model by \(RW(\epsilon, \tau, n, p)\).

Formally, our goal is to use the training data to design a classifier for use on fresh data. If we are given a new unlabeled feature vector \(X\), we must then label it with a class prediction, i.e., attach a label \(\hat{Y} = 1\) or \(\hat{Y} = -1\). We hope that our predicted label \(\hat{Y}\) is typically correct. The central problem is for which combinations \((\epsilon, \tau, n, p)\) is it possible to train a classifier that can label \(Y\) correctly, and for which combinations is it not possible to do so?

Linking Rarity and Weakness to Number of Features.

We now adopt an asymptotic viewpoint. We let the number of features \(p\) be the driving problem size descriptor, and for the purposes of calculation, we let \(p\) tend to infinity, and other quantities vary with \(p\). We have checked that our asymptotic calculations are descriptive of actual classifier performance in realistic finite-sized problems, say...
with $p$ in the few thousands, as is now common in genomics and proteomics. Other problem parameters (fraction and strength of useful features, sample size $n$) will depend on $p$ as follows. Fixing parameters $(\beta, r) \in (0, 1)^2$, let

$$
\epsilon = \epsilon_p = p^{-\beta}, \quad \tau = \tau_p = \sqrt{2\log p}.
$$

As $p \to \infty$, the useful features become increasingly rare; an asymptotically negligible fraction of the components in the vector $Z$. The parameters $(\beta, r)$ describe the linkage between rareness and weakness of the entries in the parameter vector; they have been used before in classification studies (5) and more generally in detection studies (7–9). The domain $(\beta, r) \in (0, 1)^2$ has been shown in earlier work to have an interesting two-phase structure; one can show there is a curve such that certain procedures succeed asymptotically when $(\beta, r)$ lies above the curve and fail when $(\beta, r)$ lies below the curve. We call a depiction of this domain and its phases a phase diagram. In this article, we exhibit a phase diagram such that, in the failure phase, every sequence of classification rules must fail for large $p$.

**Linking Number of Observations to Number of Features.** In classical statistical theory, one held $p$ fixed and let $n$ increase indefinitely. However, in modern scientific practice it seems the reverse is happening: one forms the impression that $n$ stays fixed or grows very weakly while $p$ grows dramatically (as high-throughput devices measure ever more features).

The phase diagram depends on the relationship between the number of features $p$ and the number of study units $n$. Again, in our work it is convenient to make $p$ the driving variable, and so $n = n_p$. We can identify three regimes for the linkage between $n$ and $p$: $n = n_p$ can have no growth, slow growth, or regular growth. Our labels for these regimes and their definitions are listed in Table 1.

The case of slow growth in our setting was previously studied in ref. 5; there, the focus was on the performance of a specific classifier-training procedure. Here, we study several types of linkages between $n$ and $p$, and we also briefly discuss the case where $n$ has an irregular growth, (see below). We are interested in limits that classifier-training procedures must obey.

**Asymptotic Rare/Weak Model (ARW).** Combining the two linkages we have just discussed gives us the asymptotic rare/weak model $ARW(\beta, r, n_p)$. For each linkage type $n_p$, we seek to identify ranges of $(\beta, r)$ where successful classification is possible and impossible, respectively.

**Impossibility of Classification.** We will show that in each of the three growth regimes, there is a curse $r = r^*(\beta) \in (N, S, R)$ which partitions the $\beta-r$ plane into two components: a region of impossibility below the curve and a region of possibility above it. In detail, define the standard phase boundary function

$$
\rho(\beta) = \begin{cases} 
0, & 0 < \beta < 1, \\
\beta - 1/2, & 1/2 < \beta < 3/4, \\
(1 - \sqrt{1 - \beta})^2, & 3/4 < \beta < 1.
\end{cases}
$$

The function $\rho$ has appeared before in determining phase boundaries in a seemingly unrelated problem of multiple hypothesis testing (7–9). Define $\rho^N(\beta) = \rho^N(\beta, n) = n^{-1} \rho(\beta)$ and $\rho^S(\beta) = \rho(\beta)$ for $0 < \beta < 1$;

$$
\rho^N(\beta) = \rho^N(\beta, n) = \frac{n^{1/2}}{n+1}\rho(\beta), \quad 0 < \beta < 1,
$$

$$
\rho^S(\beta) = \rho(\beta), \quad 0 < \beta < 1,
$$

$$
\rho^S(\beta) = (1 - \theta)\rho(\beta/(1 - \theta)), \quad 0 < \beta < (1 - \theta).
$$

See Fig. 1. Note that in the definition of $\rho^S(\beta)$, we limit $\beta$ to the range $(0, 1 - \theta)$. Note also that for all three cases,

$$
\rho^S(\beta) \leq \beta, \quad \star = N, S, R.
$$

**Definition 1:** Fix $(\beta, r) \in (0, 1)^2$. Fix one of the three types of growth of $n$ by choosing $\star \in \{N, S, R\}$. We say $(\beta, r)$ fall in the region of impossibility of the ARW if $r < \rho^\star(\beta)$.

**Theorem 1:** Fix a growth regime $n_p$, and fix a point $(\beta, r)$ in the region “below” the corresponding graph $(\beta, r^\star(\beta))$. Consider the sequence of problems $ARW(r, \beta, n_p)$ for increasing $p$ and a sequence of classifier-training methods, perhaps also dependent on $p$. The misclassification error rate of the resulting sequence of trained classifiers $\to 1/2$ with increasing $p$.

In this region, the measurements are effectively noninformative, and random guessing does almost as well. However, note that there are useful features among the $p$ features, and if we only knew which features they were, we could reliably separate the classes! Indeed, simply summing the coordinates known to be useful features and taking the signum would do the trick. In this sense, the region of impossibility is precisely the region where the effect mentioned in the introduction shows up; the attempt to find the useful features among many useless ones is simply doomed. Note that Fan and Fan (10) studied a closely related setting and identified a different region of impossibility. See details therein.

Conversely, one can show that fixing $(\beta, r)$ in the region of possibility, successful classifier training is possible, and there is a sequence of trained classifiers whose misclassification probability $\to 0$ as $p \to \infty$. However, that is beyond the scope of this research announcement; we refer the reader to the author’s related papers. See Fig. 2 for a display of region of impossibility and region of possibility.

**Proof of Theorem 1:** To understand the role of the training data, we compare the problem of classification with $Z$ and that without $Z$. When $Z$ is not available, the test features $X$ contains $p_\rho$ useful features, each of which has a strength of $\pm \tau_p/\sqrt{n}$ (with sign “+” if

### Table 1. Regimes and their definitions

<table>
<thead>
<tr>
<th>Regime</th>
<th>Label</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>No growth</td>
<td>(N)</td>
<td>$n_p = n_0$ for some constant $n_0$</td>
</tr>
<tr>
<td>Slow growth</td>
<td>(S)</td>
<td>$n_p \to \infty$, but $n_p/p^\rho \to 0$, $\forall \theta &gt; 0$</td>
</tr>
<tr>
<td>Regular growth</td>
<td>(R)</td>
<td>$n_p = p^\theta$ for some $\theta \in (0, 1)$</td>
</tr>
</tbody>
</table>
\(X\) is from class 1 and “−” otherwise). In this case, the classification problem reduces to the testing problem studied in our previous work \((8)\), and it is possible to successfully classify if and only if \(r/s > \rho(\beta)\) \((7, 8)\).

When \(Z\) is available to us, the picture is very different. For any \(1 \leq i \leq p\), the probability that the \(i\)th coordinate of \(X\) contains a useful feature is no longer \(\epsilon_p\), but instead the posterior probability \(\eta_i = \eta(Z_i; p)\), where

\[\eta(z; p) = \epsilon_p \psi(z - \tau_p)/(1 - \epsilon_p) \psi(z) + \epsilon_p \psi(z - \tau_p),\]

with \(\psi\) being the density of \(N(0, 1)\). This is a monotone function, \(\approx 0\) for small \(z\) and \(\approx 1\) for large \(z\). Intuitively, a large coordinate amplitude in \(Z\) suggests a useful feature, and a moderate or small amplitude suggests a useless one. Seemingly, this is a different model from the previous case, the study of which needs different analytic technique.

First, denote the density of \(N(0, I_p)\) by \(f_0^{(p)} = f_0^{(p)}(x_1, x_2, \ldots, x_p)\). Second, for \(k = 1, 2\), denote the conditional density of \((X|Z)\) by

\[f_k^{(p)} = f_k^{(p)}(x_1, x_2, \ldots, x_p|Z; \epsilon_p, \tau_p, \eta_p, p),\]

and denote the conditional density of \((X_i|Z_i)\) by

\[f_k^{(1)} = f_k^{(1)}(x_1|Z; \epsilon_p, \tau_p, \eta_p).\]

Here, \(X_i\) and \(Z_i\) are the first coordinates of \(X\) and \(Z\), respectively. Finally, for two density functions \(f\) and \(g\), define the Hellinger affinity by \(H(f, g) = \int \sqrt{f(x)g(x)}dx\). Let the (conditional)-Hellinger affinity between \(f_0^{(p)}\) and \(f_1^{(p)}\) be \(H_0^{(p)}(f_0^{(p)}; f_1^{(p)}; Z; \epsilon_p, \tau_p, \eta_p, p)\), and that between \(f_0^{(1)}\) and \(f_1^{(1)}\) be \(H_0^{(1)}(f_0^{(1)}; f_1^{(1)}; Z; \epsilon_p, \tau_p, \eta_p, p)\). We have the following lemma.

**Lemma 1.** Fix \(n_p\) and \((\beta, r) \in (0, 1)^2\) in the ARW \((\beta, r, n_p)\) model. For any classifier \(T = T(X, Z; p)\)

\[
\begin{align*}
&P[\text{misclassification}\mid T] = 1/2] \\
&\leq C(1 - E[H_0^{(p)}(f_0^{(p)}; f_1^{(p)}; Z; \epsilon_p, \tau_p, \eta_p, p)]^{1/2}.
\end{align*}
\]

We omit the proof of **Lemma 1**. Relationships between classification error rate and Hellinger affinity are well known. In this case, the added wrinkle is condition on the training data \(Z\). Besides that feature, the argument is standard; see ref. 11 for example.

The following lemma is elementary; we omit the proof.

**Lemma 2.** \(E[H_0^{(p)}(f_0^{(1)}; f_1^{(1)}; Z; \epsilon_p, \tau_p, \eta_p, p)] = (E[H_0^{(1)}(f_0^{(1)}; f_1^{(1)}; Z; \epsilon_p, \tau_p, \eta_p, p)])^p.\)

The heart of the proof of **Theorem 1** is the following lemma. Its proof is relatively long, so we leave it to later sections.

**Lemma 3.** Fix one of the three growth types; for any fixed parameters \((\beta, r)\) in the corresponding region of impossibility of the ARW \((\beta, r, n_p)\),

\[E[H_0^{(p)}(f_0^{(1)}; f_1^{(1)}; Z; \epsilon_p, \tau_p, \eta_p, p)] = 1 + o(1/p),\]

\(p \rightarrow \infty.\)

Combining these lemmas gives **Theorem 1**.

**Extension to Cases Where \(n\) Grows Irregularly.** So far, we considered \(n_p\) growing with \(p\) according to one of three specific regimes: (N), (S), (R). However, the conclusion of **Theorem 1** can be obtained in a much broader range of cases, where \(n\) grows somewhat irregularly.

**Lemma 4** below says that \(E[H_0^{(p)}(f_0^{(1)}; f_1^{(1)}; Z; \epsilon, \tau, n, p)]\) is a monotone function of the sample size \(n\). This implies that, if \(n_p\) is eventually sandwiched between two sequences obeying one of our growth regimes, then its behavior is also sandwiched between the results at those two regular situations.

**Lemma 4.** Fixing \(p, \tau > 0\) and \(\epsilon \in (0, 1)\) in the RW \((\epsilon, \tau, n, p)\), \(E[H_0^{(p)}(f_0^{(1)}; f_1^{(1)}; Z; \epsilon, \tau, n, p)]\) is a monotone increasing function of \(n\).

**Proof.** By **Lemma 2**, it suffices to show that \(E[H_0^{(1)}(f_0^{(1)}; f_1^{(1)}; Z; \epsilon, \tau, n, p)]\) is a monotone increasing function of \(n\). Note that \(f_0^{(1)}\) is the density of \(N(0, 1)\), and that \(f_0^{(1)}(x) = (1 - \eta(Z_i))\psi(x) + \eta(Z_i)\psi(x - \mu_p),\) with \(\eta(Z_i) = \eta(Z_i; p)\) being defined in Eq. 3, and \(\mu_p = \tau_p/\sqrt{n}\). By direct calculations,

\[
H_0^{(1)}(f_0^{(1)}; f_1^{(1)}; Z, 1) = E \left[1 - \eta(Z_i) + \eta(Z_i)e^{\mu_p X_i - \mu^2/2} \right]^{1/2},
\]

where we have suppressed parameter dependencies and \(E_0\) denotes the expectation with respect to the law of \(X \sim N(0, 1)\). Observe that \(H_0^{(1)}(f_0^{(1)}; f_1^{(1)}; Z, 1; \epsilon, \tau, n, p)\) depends on \(n\) only through \(\mu_p\), and that \(\mu_p\) is monotone decreasing in \(n\). It is sufficient to show that for any \(\ell \in (0, 1)\) and \(\mu^2 \geq \mu\),

\[
E_0 \left[\sqrt{1 - \eta} + \eta e^{\mu X - \mu^2/2} \right] \leq E_0 \left[\sqrt{1 - \eta} + \eta e^{\mu X - \mu^2/2} \right].
\]

**[5]**

Toward this end, write \(\mu X_i = \mu U + \delta W\), where \(U\) and \(W\) are iid samples from \(N(0, 1)\) and \(\delta^2 = (\mu^2)^2 - \mu^2\). Inserting this into the left-hand side of Eq. 5 gives

\[
E_0 \left[\sqrt{1 - \eta} + \eta e^{\mu X - \mu^2/2} \right] = E_0 \left[\sqrt{1 - \eta} + \eta e^{(\mu^2 - \mu^2)/2} \right].
\]

It follows from Jensen’s inequality that

\[
E_0 \left[\sqrt{1 - \eta} + \eta e^{(\mu^2 - \mu^2)/2} \right] \leq E_0 \left[\sqrt{1 - \eta} + \eta e^{(\mu^2 - \mu^2)/2} \right].
\]
where the right-hand side reduces to that of Eq. 5. This gives the claim.

**Gaussian Assumption.** We now discuss the Gaussian assumption on $Z$ (that on $X_i$ has a less important effect). When $n$ is relatively large (e.g., $n \geq 20$), the assumption is reasonable. When $n$ is very small, it might be better to assume that $Z(j)$ are $t$-distributed. Because the tail of the $t$ distribution is heavier than that of the Gaussian, it is harder to successfully classify in the $t$-error model than in the Gaussian model, given that the parameters ($p,n,\epsilon, \tau$) are the same in two settings. Therefore, the region of impossibility continues to be valid in the $t$ setting. Of course, the exact boundary separating region of possibility and region of impossibility depends on the specific tail behavior of the marginal density of $Z(j)$, and may be different from that in this note.

**Numerical Examples.** In companion articles (4, 5), we show that if we perform ideal feature selection with an oracle threshold—i.e., if an oracle tells us the unknown parameters ($\beta, r$)—then the resulting trained classifier yields successful classification throughout the region of possibility. Second, we show that when the HCT converges to the oracle threshold asymptotically (but does not require help from any oracle).

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>$\epsilon$</th>
<th>$\tau$</th>
<th># of misclassifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.55</td>
<td>0.45</td>
<td>1.25</td>
<td>0.001</td>
</tr>
<tr>
<td>0.65</td>
<td>0.35</td>
<td>1.55</td>
<td>0.002</td>
</tr>
<tr>
<td>0.75</td>
<td>0.25</td>
<td>1.85</td>
<td>0.003</td>
</tr>
<tr>
<td>0.85</td>
<td>0.15</td>
<td>2.15</td>
<td>0.004</td>
</tr>
</tbody>
</table>

The HCT statistic is the maximum of the objective function. It can be used to assess significance of the whole body of $P$ values. Given the feature vector $X$ as in the ARW (but not any class labels), test whether the mean vector $\mu = 0$ identically, or whether $\mu$ contains a fraction of nonzero coordinates, of each of them equal to an unknown parameter $\tau_s$. The testing problem is a modification of the classification problem we study in this note, where the training set is not available. Similarly, the testing problem was shown in refs. 7 and 8 to have a phase diagram ($\beta, r$) with a region of impossibility and a region of possibility. In fact, if we express $\epsilon = 0.5$ for the critical value is arbitrary; other choices produce quantitatively similar but not identical results).

When $n = 2$ (and $\beta$ takes corresponding values as above), $\tau_s \approx 0.9/1.5/2.0$; when $n = 10$, $\tau_s \approx 1.3/1.9/2.4$; when $n = 100$, $\tau_s \approx 0.7/1.3/1.9$.

**Table 2. Misclassification errors for $L(X; t)$ with ideal thresholds ($p = 10^5$) ($\epsilon = p^{-\beta}$)**

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\tau$</th>
<th>$\beta$</th>
<th># of misclassifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.55</td>
<td>0.45</td>
<td>0.001</td>
</tr>
<tr>
<td>10</td>
<td>0.65</td>
<td>0.35</td>
<td>0.002</td>
</tr>
<tr>
<td>100</td>
<td>0.75</td>
<td>0.25</td>
<td>0.003</td>
</tr>
<tr>
<td>500</td>
<td>0.85</td>
<td>0.15</td>
<td>0.004</td>
</tr>
</tbody>
</table>

The HC statistic is the maximum of the objective function. It can be used to assess significance of the whole body of $P$ values. Given the feature vector $X$ as in the ARW (but not any class labels), test whether the mean vector $\mu = 0$ identically, or whether $\mu$ contains a fraction of nonzero coordinates, of each of them equal to an unknown parameter $\tau_s$. The testing problem is a modification of the classification problem we study in this note, where the training set is not available. Similarly, the testing problem was shown in refs. 7 and 8 to have a phase diagram ($\beta, r$) with a region of impossibility and a region of possibility. In fact, if we express $\epsilon = 0.5$ for the critical value is arbitrary; other choices produce quantitatively similar but not identical results).

When $n = 2$ (and $\beta$ takes corresponding values as above), $\tau_s \approx 0.9/1.5/2.0$; when $n = 10$, $\tau_s \approx 1.3/1.9/2.4$; when $n = 100$, $\tau_s \approx 0.7/1.3/1.9$.

**Table 2. Misclassification errors for $L(X; t)$ with ideal thresholds ($p = 10^5$) ($\epsilon = p^{-\beta}$)**

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\tau$</th>
<th>$\beta$</th>
<th># of misclassifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.55</td>
<td>0.45</td>
<td>0.001</td>
</tr>
<tr>
<td>10</td>
<td>0.65</td>
<td>0.35</td>
<td>0.002</td>
</tr>
<tr>
<td>100</td>
<td>0.75</td>
<td>0.25</td>
<td>0.003</td>
</tr>
<tr>
<td>500</td>
<td>0.85</td>
<td>0.15</td>
<td>0.004</td>
</tr>
</tbody>
</table>
In this article, we have attempted to draw the attention of working scientists to a basic phenomenon that might affect many high-throughput studies. The mathematical scientist interested in this phenomenon will want to know that independently, Ingster, Pouet, and Tsybakov (15) have analyzed a setting more general than the present one and identified similar phase transitions phenomena.

**Proof of Lemma 3:** We now show Lemma 3 for the case of no growth and the case of regular growth. Once these are proved, the case of slow growth follows by the monotonicity result in Lemma 4 and the way $r^*(p)$ is defined.

The case of no growth $(N)$. In this case, $n$ is a fixed integer. It suffices to show for fixed $(\beta, r)$ with $r < \frac{\beta - 1}{\beta + 1}p(\beta)$,

$$E[H(f_0^{(1)} f_1^{(1)}; Z_1)] = 1 + O(1/p).$$

Note that in the ARW, the density of $Z_1$ is $(1 - \epsilon_p)\psi(x) + \epsilon_p\psi(x - \tau_p)$.

By Eq. 4 and direct calculations,

$$E[H(f_0^{(1)} f_1^{(1)}; Z_1)] = (1 - \epsilon_p)H + \epsilon_pH,$$

where

$$I = E[0, \frac{1}{\epsilon_p} + \epsilon_p \psi(x - \frac{1}{\epsilon_p})],$$

$$H = E[0, \frac{1}{\epsilon_p} + \epsilon_p \psi(x - \frac{1}{\epsilon_p})],$$

and $E[0, \frac{1}{\epsilon_p} + \epsilon_p \psi(x - \frac{1}{\epsilon_p})]$ is the expectation of the law with respect to $X$ and $Z$ are iid samples from $N(0, 1)$. Write for short $E_0 = E[0, \frac{1}{\epsilon_p} + \epsilon_p \psi(x - \frac{1}{\epsilon_p})]$ whenever there is no confusion. Introduce $\epsilon_p = 1/(1 - \epsilon_p)$.

As $X$ and $Z$ are iid samples from $N(0, 1), so W \sim N(0, 1). Using Lemma 5,

$$E_0[V_1(\epsilon_p e^{\tau_p Z - \frac{1}{\epsilon_p}}, e^{\psi(x - \frac{1}{\epsilon_p})})]$$

$$\leq C E_0[\psi_1(\epsilon_p e^{\tau_p Z - \frac{1}{\epsilon_p}}, e^{\psi(x - \frac{1}{\epsilon_p})})]$$

$$+ E_0[(1 + e^{\psi(x - \frac{1}{\epsilon_p})}) \psi_1(\epsilon_p e^{\tau_p Z - \frac{1}{\epsilon_p}})]$$

$$= C E_0[(\psi_1(\epsilon_p e^{\tau_p Z - \frac{1}{\epsilon_p}})] + E_0(1 + e^{\psi(x - \frac{1}{\epsilon_p})}) \psi_1(\epsilon_p e^{\tau_p Z - \frac{1}{\epsilon_p}})]$$

$$\leq C E_0[(\psi_1(\epsilon_p e^{\tau_p Z - \frac{1}{\epsilon_p}})]$$

The second term in (16) is no greater than the first term. To see this, writing $\psi_1(\epsilon_p e^{\tau_p Z - \frac{1}{\epsilon_p}}) = \psi_2(\epsilon_p e^{\tau_p Z - \frac{1}{\epsilon_p}}, e^{\psi(x - \frac{1}{\epsilon_p})})$, it follows from Eq. 8, the convexity of $\psi_1$, and Jensen’s inequality, that

$$E_0[\psi_1(\epsilon_p e^{\tau_p Z - \frac{1}{\epsilon_p}})] = E_0[(\psi_1(\epsilon_p e^{\tau_p Z - \frac{1}{\epsilon_p}})]$$

$$\leq CE_0[(\psi_1(\epsilon_p e^{\tau_p Z - \frac{1}{\epsilon_p}})]$$

which validates the aforementioned point. Combining this with Eq. 15 gives

$$E_0[\psi_1(\epsilon_p e^{\tau_p Z - \frac{1}{\epsilon_p}})] + E_0[(1 + e^{\psi(x - \frac{1}{\epsilon_p})}) \psi_1(\epsilon_p e^{\tau_p Z - \frac{1}{\epsilon_p}})]$$

$$\leq CE_0[(\psi_1(\epsilon_p e^{\tau_p Z - \frac{1}{\epsilon_p}})]$$

We now analyze $E_0[\psi_1(\epsilon_p e^{\tau_p Z - \frac{1}{\epsilon_p}})]$. Introduce $\rho_0 = \frac{\beta - 1}{\beta + 1}p(\beta)$. Recall that $\psi_1(\epsilon_p e^{\tau_p Z - \frac{1}{\epsilon_p}})$.

By these definitions and basic algebra,

$$\epsilon_p e^{\tau_p Z - \frac{1}{\epsilon_p}} > 1$$

Combining Eq. 18 with the definition of $\psi_1$,

$$E_0[\psi_1(\epsilon_p e^{\tau_p Z - \frac{1}{\epsilon_p}})]$$

$$\leq 2 \int_0^{\epsilon_p} e^{\psi(x - \frac{1}{\epsilon_p})} \psi(w)dw + \int_{\epsilon_p}^{\frac{\beta}{\beta + 1}p(\beta)} e^{\psi(x - \frac{1}{\epsilon_p})} \psi(w)dw$$

$$= 2 \epsilon_p \psi - (\epsilon_p - \sigma_p) + \epsilon_p^2 \psi^2 \phi(p_0 - 2\sigma_p),$$

where $\phi$ is the cdf of $N(0, 1)$. In addition, by the assumption $r < \frac{\beta - 1}{\beta + 1}p(\beta)$, we have $r_0 < \rho(\beta)$. In view of the definitions of $t_p, \sigma_p$, and $\rho(\beta)$, it follows from basic algebra that

$$1 < \frac{t_p}{\sigma_p} \leq 2 \text{ if } r_0 \geq \beta/3; \quad \frac{t_p}{\sigma_p} > 2 \text{ if } r_0 > \beta/3.$$
Recall that \( r_0 < \rho(\beta) \). It follows from the definition of \( \rho(\beta) \) that 
\[
\frac{(\beta r)^2}{\alpha_0} > 1, \text{ and that } 2\beta > 2\theta_0 > 1.
\]
As a result, 
\[
E_0 \left[ \psi_2(\epsilon \phi^2 pX - \psi^2 / 2) \right] = o(1/p). \tag{22}
\]

Inserting Eq. 22 into Eq. 16 gives Eq. 11. Next, consider Eq. 12. Similarly, by Lemma 5, Eq. 8, and that X and Z are independent, 
\[
E_0 \left[ V_2 \left( \epsilon \phi^2 pX + \psi^2 / 2, \epsilon \phi^2 pX - \psi^2 / 2 \right) \right] \leq CE_0 \left[ \left( 1 + \epsilon \phi^2 pX - \psi^2 / 2 \right) \psi_2(\epsilon \phi^2 pZ + \psi^2 / 2) \right]. \tag{23}
\]

Similarly, as that \( \psi_2 e^{\phi^2 pZ + \psi^2 / 2} > 1 \) if and only if \( Z > t_p - \tau_p \), by the definition of \( \psi_2 \) and elementary calculus,
\[
E_0 \left[ \psi_2 \left( \epsilon \phi^2 pZ + \psi^2 / 2 \right) \right] \leq \int_{-\infty}^{\infty} \psi_2(z)dz + \int_{t_p - \tau_p}^{\infty} \epsilon \phi^2 pZ + \psi^2 / 2 \psi(z)dz = \Phi(-t_p + \tau_p) + \epsilon \phi^2 \Phi(t_p - 2\tau_p). \tag{24}
\]

Combine Eq. 25 with Eqs. 20 and 21,
\[
E_0 \left[ \psi_2(\epsilon \phi^2 pZ + \psi^2 / 2) \right] \leq \left\{ \begin{array}{ll}
C p \left( \frac{\beta - r^2}{\omega} \right), & r \geq \beta/3, \\
C \left( \frac{p - \beta}{\omega} + p(\beta - r^2) \right), & r < \beta /3.
\end{array} \right. \tag{26}
\]

Now, since \( r < \rho(\beta) \), then \( \left( \frac{\beta - r^2}{\omega} \right) > (1 - \beta) \) and \( (\beta - 2r) > 1 - \beta \). Inserting these into Eq. 26 gives Eq. 12, and concludes the proof for the case of no growth.

The case of regular growth (R). By Eq. 2, we can limit \( (\beta, r) \) to the range of \( 0 < r < \beta \) and \( 0 < \beta < 1 \). Define
\[
\delta(r, \beta) = \left\{ \begin{array}{ll}
\frac{r - \beta}{\beta}, & \text{if } r \leq \beta /3, \\
\left( \frac{r - \beta}{\beta} - \frac{(\beta - 3^2)}{\omega} \right), & \text{if } r > \beta /3.
\end{array} \right. \tag{27}
\]

Basic algebra shows that the assumption \( r < (1 - \theta \rho(\beta)/(1 - \theta)) \) is equivalent to \( 2\delta(r, \beta) < \theta \). Recall that the sample size \( n = p^2 \). It is sufficient to show that for fixed \( (\beta, r, \theta) \) satisfying \( 2\delta(r, \beta) < \theta \), 
\[
E[H(f_0, f_1; Z)] = 1 + o(1/p). \tag{28}
\]

Rewrite \( H(f_0, f_1; Z) = E_0[1 + \eta_0(Z)(e^{\phi^2 pX - \psi^2 / 2} - 1)^2] \). Since \( |\sqrt{1+x - 1 - x^2} / \omega | \leq C x^2 \) for any \( x > -1 \), then
\[
\left| H(f_0, f_1; Z) - 1 - \frac{1}{2} \eta_0(Z) E_0 \left[ e^{\phi^2 pX - \psi^2 / 2} - 1 \right] \right| \leq C \eta_0^2(Z) E_0 \left[ \left( e^{\phi^2 pX - \psi^2 / 2} - 1 \right)^2 \right]. \tag{29}
\]

Recalling \( n = p^2 \) and \( \mu_p = \tau_p / \sqrt{n} \), direct calculations show that
\[
E_0 \left[ e^{\phi^2 pX - \psi^2 / 2} - 1 \right] = 0, \tag{30}
\]
\[
E_0 \left[ e^{\phi^2 pX - \psi^2 / 2} - 1 \right]^2 = \epsilon^2 p^2 \leq 1 \leq C \log(p)p^{-\theta}, \tag{31}
\]
Combining Eqs. 30 and 31 with Eq. 29 gives \( |H(f_0, f_1; Z) - 1| \leq C \log(p)p^{-\theta} \eta_0^2(Z) \). If we can show that for any \( (\beta, r) \) in the range of \( 0 < r < \beta \) and \( 0 < \beta < 1 \), 
\[
E[\eta_0^2(Z)] \leq C p^{-2-2\theta(r, \beta)}, \tag{32}
\]
then \( |E[H(f_0, f_1; Z)] - 1| \leq E[H(f_0, f_1; Z)] \leq C \log(p)p^{-1+2\theta(r, \beta)} \), and Eq. 28 follows from the assumption of \( 2\delta(r, \beta) < \theta \). We now show Eq. 32. Write \( E[\eta_0^2(Z)] = I + II \), where 
\[
I = (1 - \epsilon_p) E_0[\eta_0^2(Z)], \quad II = \epsilon_p E_0[\eta_0^2(\tau_p + Z)]. \tag{33}
\]

With \( E_0 \) being the expectation with respect to the law of \( Z \sim N(0, 1) \). It is sufficient to show that 
\[
I \leq C p^{-1+2\theta(r, \beta)}, \quad II \leq C p^{-1+2\theta(r, \beta)}. \tag{34}
\]

Consider the first claim of Eq. 34. Recall that \( \tau_p = \sqrt{2r \log p} \) and \( t_p = (r, \beta) = \frac{1}{\sqrt{2r}} \tau_p \), and note that for sufficiently large \( p \), 
\[
(1 - \epsilon_p) + e^{\phi^2 pX - \psi^2 / 2} \geq 1/2. \tag{32}
\]

Combine this with the way that \( \eta_0(Z) \) is defined, \( \eta_0(Z) \leq 2\epsilon_p e^{\phi^2 pX - \psi^2 / 2} \) when \( Z \leq t_p \) and \( \eta_0(Z) \leq 1 \) otherwise. It follows from these inequalities and elementary calculus that 
\[
I \leq C p e^{\phi^2 \Phi(t_p - 2\tau_p)} + \Phi(-t_p). \tag{35}
\]

By arguments similar to that in the proof for the case of no growth, 
\[
\Phi(-t_p) \leq C \psi_2(t_p) \leq C \psi_2 \left( \frac{\beta - r^2}{\omega} \right), \tag{36}
\]
and 
\[
e^{\phi^2 \Phi(t_p - 2\tau_p)} \leq \left\{ \begin{array}{ll}
\frac{n^{-2r}}{C \epsilon_p \delta(r, \beta)^2}, & \text{if } r \leq \beta /3, \\
\frac{C n^{-2r} \delta^2 (r, \beta)^2}{C r^2}, & \text{if } r > \beta /3.
\end{array} \right. \tag{37}
\]

Inserting Eqs. 36 and 37 into Eq. 35 gives the claim. Consider the second claim of Eq. 34. By similar argument, 
\[
II \leq C \epsilon_p^2 e^{\phi^2 \Phi(t_p - 3\tau_p)} + \Phi(-t_p - \tau_p)). \tag{38}
\]

and 
\[
e^{\phi^2 \Phi(t_p - 3\tau_p)} \leq \left\{ \begin{array}{ll}
\frac{n^{-2r}}{C \epsilon_p \delta(r, \beta)^2}, & \text{if } r \leq \beta /5, \\
\frac{C n^{-2r} \delta^2 (r, \beta)^2}{C r^2}, & \text{if } r > \beta /5.
\end{array} \right. \tag{39}
\]

Combining these results gives the second claim of Eq. 34, and concludes the proof for the case of regular growth.

ACKNOWLEDGMENTS. I thank David Donoho and anonymous referees for numerous helpful pointers and comments in improving the manuscript, and the Newton Institute for hospitality during the program Statistical Challenges of High-Dimensional Data. This work was supported in part by National Science Foundation Grant DMS-0908613.