0.0.1 Expanding the Set

In Section ?? we saw how to resolve the issue with TV projection by relaxing to a weaker distance $\overline{TV}$. We will now study an alternate approach, based on expanding the destination set $G$ to a larger set $M$. For this approach we will need to reference the “true empirical distribution” $p_n^*$. What we mean by this is the following: Whenever $TV(p^*, \hat{p}) \leq \epsilon$, we know that $p^*$ and $\hat{p}$ are identical except for some event $E$ of probability $\epsilon$. Therefore we can sample from $\hat{p}$ as follows:

1. Draw a sample from $X \sim p^*$.
2. Check if $E$ holds; if it does, replace $X$ with a sample from the conditional distribution $\hat{p}_{|E}$.
3. Otherwise leave $X$ as-is.

Thus we can interpret a sample from $\hat{p}$ as having a $1 - \epsilon$ chance of being “from” $p^*$. More generally, we can construct the empirical distribution $\tilde{p}_n$ by first constructing the empirical distribution $p_n^*$ coming from $p^*$, then replacing $\text{Binom}(n, \epsilon)$ of the points with samples from $\tilde{p}_{|E}$. Formally, we have created a coupling between the random variables $p_n^*$ and $\tilde{p}_n$ such that $TV(p_n^*, \tilde{p}_n)$ is distributed as $\frac{1}{n} \text{Binom}(n, \epsilon)$.

Let us return to expanding the set from $G$ to $M$. For this to work, we need three properties to hold:

- $M$ is large enough: $\min_{q \in M} TV(q, p_n^*)$ is small with high probability.
- The empirical loss $L(p_n^*, \theta)$ is a good approximation to the population loss $L(p^*, \theta)$.
- The modulus is still bounded: $\min_{p, q \in M : TV(p,q) \leq 2\epsilon} L(p, \theta^*(q))$ is small.

In fact, it suffices for $M$ to satisfy a weaker property; we only need the “generalized modulus” to be small relative to some $G' \subset M$:

**Proposition 0.1.** For a set $G' \subset M$, define the generalized modulus of continuity as

$$m(G', M, 2\epsilon) \overset{\text{def}}{=} \min_{p \in G', q \in M : TV(p,q) \leq 2\epsilon} L(p, \theta^*(q)).$$

Assume that the true empirical distribution $p_n^*$ lies in $G'$ with probability $1 - \delta$. Then the minimum distance functional projecting under $TV$ onto $M$ has empirical error $L(p_n^*, \hat{\theta})$ at most $m(G', M, 2\epsilon)$ with probability at least $1 - \delta - \mathbb{P}[\text{Binom}(\epsilon, n) \geq \epsilon n]$.

**Proof.** Let $\epsilon' = TV(p_n^*, \tilde{p}_n)$, which is $\text{Binom}(\epsilon, n)$-distributed. If $p_n^*$ lies in $G'$, then since $G' \subset M$ we know that $\tilde{p}_n$ has distance at most $\epsilon'$ from $M$, and so the projected distribution $q$ satisfies $TV(q, \tilde{p}_n) \leq \epsilon'$ and hence $TV(q, p_n^*) \leq 2\epsilon'$. It follows from the definition that $L(p_n^*, \hat{\theta}) = L(p_n^*, \theta^*(q)) \leq m(G', M, 2\epsilon')$. □

A useful bound on the binomial tail is that $\mathbb{P}[\text{Binom}(\epsilon, n) \geq 2\epsilon n] \leq \exp(-\epsilon n / 3)$. In particular the empirical error is at most $m(G', M, 4\epsilon)$ with probability at least $1 - \delta - \exp(-\epsilon n / 3)$.

**Application: bounded kth moments.** First suppose that the distribution $p^*$ has bounded $k$th moments, i.e. $G_{\text{mom}, k}(\sigma) = \{ p | \| p \|_{\psi} \leq \sigma \}$, where $\psi(x) = x^k$. When $k > 2$, the empirical distribution $p_n^*$ will not have bounded $k$th moments until $n \geq \Omega(d^{k/2})$. This is because if we take a sample $x_1 \sim p$ and let $v$ be a unit vector in the direction of $x_1 - \mu$, then $E_{x \sim p_n^*}[|x - \mu, v|^k] \geq \frac{1}{n} \| x_1 - \mu \|^k \geq d^{k/2} / n$, since the norm of $\| x_1 - \mu \|^2$ is typically $\sqrt{d}$.

Consequently, it is necessary to expand the set and we will choose $G' = M = G_{TV}(\rho, \epsilon)$ for $\rho = O(\sigma \epsilon^{-1/k})$ to be the set of resilience distributions with appropriate parameters $\rho$ and $\epsilon$. We already know that the modulus of $M$ is bounded by $O(\sigma \epsilon^{-1/k})$, so the hard part is showing that the empirical distribution $p_n^*$ lies in $M$ with high probability.

As noted above, we cannot hope to prove that $p_n^*$ has bounded moments except when $n = \Omega(d^{k/2})$, which is too large. We will instead show that certain truncated moments of $p_n^*$ are bounded as soon as $n = \Omega(d)$,
and that these truncated moments suffice to show resilience. Specifically, if \( \psi(x) = x^k \) is the Orlicz function for the \( k \)th moments, we will define the truncated function

\[
\tilde{\psi}(x) = \begin{cases} 
    x^k : x \leq x_0 \\
    kx_0^{k-1}(x - x_0) + x_0^k : x > x_0
\end{cases}
\]

(2)

In other words, \( \tilde{\psi} \) is equal to \( \psi \) for \( x \leq x_0 \), and is the best linear lower bound to \( \psi \) for \( x > x_0 \). Note that \( \tilde{\psi} \) is \( L \)-Lipschitz for \( L = kx_0^{k-1} \). We will eventually take \( x_0 = (k^{k-1}\epsilon^{-1/2})^{-1/k} \) and hence \( L = (1/\epsilon)^{(k-1)/k} \). Using a symmetrization argument, we will bound the truncated \( \sup_{\|v\|_2 \leq 1} \mathbb{E}_{P^*}[\psi(\|x - \mu, v\|/\sigma)] \).

**Proposition 0.2.** Let \( X_1, \ldots, X_n \sim p^* \), where \( p^* \in \mathcal{G}_{\text{mom}, k}(\sigma) \). Then,

\[
\mathbb{E}_{X_1, \ldots, X_n \sim p^*} \left[ \sup_{\|v\|_2 \leq 1} \frac{1}{n} \sum_{i=1}^n \tilde{\psi}(\frac{\|X_i - \mu, v\|}{\sigma}) - U(v) \right] \leq O\left( 2L\sqrt{\frac{dk}{n}} \right),
\]

(3)

where \( L = kx_0^{k-1} \) and \( U(v) \) is a function satisfying \( U(v) \leq 1 \) for all \( v \).

Before proving Proposition 0.2, let us interpret its significance. Take \( x_0 = (k^{k-1}\epsilon^{-1/2})^{-1/k} \) and hence \( L = \epsilon^{-1/k} \). Take \( n \) large enough so that the right-hand-side of (3) is at most 1, which requires \( n \geq \Omega(kd/\epsilon^{2-2/k}) \). We then obtain a high-probability bound on the \( \psi \)-norm of \( p_n^* \), i.e. the \( \tilde{\psi} \)-norm is at most \( O(\delta^{-1/k}) \) with probability \( 1 - \delta \). This implies that \( p_n^* \) is resilient with parameter \( \rho = \sigma\epsilon \tilde{\psi}^{-1}(\mathcal{O}(\delta^{-1/k})/\epsilon) = 2\sigma\epsilon^{1-1/k} \). A useful bound on \( \tilde{\psi}^{-1} \) is \( \tilde{\psi}^{-1}(\epsilon) \leq x_0 + z/L \), and since \( x_0 \leq (1/\epsilon)^{-1/k} \) and \( L = (1/\epsilon)^{(k-1)/k} \) in our case, we have

\[
\rho \leq O(\sigma\epsilon^{1-1/k}\delta^{-1/k}) \text{ with probability } 1 - \delta.
\]

This matches the population-bound of \( O(\sigma\epsilon^{1-1/k}) \), and only requires \( kd/\epsilon^{2-2/k} \) samples, in contrast to the \( d/\epsilon^2 \) samples required before. Indeed, this sample complexity dependence is optimal (other than the factor of \( k \)); the only drawback is that we do not get exponential tails (we instead obtain tails of \( \delta^{-1/k} \), which is worse than the \( \sqrt{\log(1/\delta)} \) from before).

Now we discuss some ideas that are needed in the proof. We would like to somehow exploit the fact that \( \tilde{\psi} \) is \( L \)-Lipschitz to prove concentration. We can do so with the following keystone result in probability theory:

**Theorem 0.3 (Ledoux-Talagrand Contraction).** Let \( \phi : \mathbb{R} \to \mathbb{R} \) be an \( L \)-Lipschitz function such that \( \phi(0) = 0 \). Then for any convex, increasing function \( g \) and Rademacher variables \( \epsilon_i \sim \{\pm 1\} \), we have

\[
\mathbb{E}_{\epsilon_i \sim \{\pm 1\}}[g(\sup_{t \in T} \sum_{i=1}^n \epsilon_i \phi(t_i))] \leq \mathbb{E}_{\epsilon_i \sim \{\pm 1\}}[g(L \sup_{t \in T} \sum_{i=1}^n \epsilon_i t_i)]
\]

(4)

Let us interpret this result. We should think of the \( t_i \) as a quantity such as \( (x_i - \mu, v) \), where abstracting to \( t_i \) yields generality and notational simplicity. Theorem 0.3 says that if we let \( Y = \sup_{t \in T} \sum_{i} \epsilon_i \phi(t_i) \) and \( Z = L \sup_{t \in T} \sum_{i} \epsilon_i t_i \), then \( \mathbb{E}[g(Y)] \leq \mathbb{E}[g(Z)] \) for all convex increasing functions \( g \). When this holds we say that \( Y \) stochastically dominates \( Z \) in second order; intuitively, it is equivalent to saying that \( Z \) has larger mean than \( Y \) and greater variation around its mean. For distributions supported on just two points, we can formalize this as follows:

**Lemma 0.4 (Two-point stochastic dominance).** Let \( Y \) take values \( y_1 \) and \( y_2 \) with probability \( \frac{1}{2} \), and \( Z \) take values \( z_1 \) and \( z_2 \) with probability \( \frac{1}{2} \). Then \( Z \) stochastically dominates \( Y \) (in second order) if and only if

\[
\frac{z_1 + z_2}{2} \geq \frac{y_1 + y_2}{2} \text{ and } \max(z_1, z_2) \geq \max(y_1, y_2).
\]

(5)

**Proof.** Without loss of generality assume \( z_2 \geq z_1 \) and \( y_2 \geq y_1 \). We want to show that \( \mathbb{E}[g(Y)] \leq \mathbb{E}[g(Z)] \) for all convex increasing \( g \) if and only if (5) holds. We first establish necessity of (5). Take \( g(x) = x \), then we require \( \mathbb{E}[Y] \leq \mathbb{E}[Z] \), which is the first condition in (5). Taking \( g(x) = \max(x - z_2, 0) \) yields \( \mathbb{E}[g(Z)] = 0 \) and \( \mathbb{E}[g(Y)] \geq \frac{1}{2} \max(y_2 - z_2, 0) \), so \( \mathbb{E}[g(Y)] \leq \mathbb{E}[g(Z)] \) implies that \( y_2 \leq z_2 \), which is the second condition in (5).
We next establish sufficiency, by conjuring up appropriate weights for Jensen’s inequality. We have
\[
\frac{y_2 - z_1}{z_2 - z_1} g(z_2) + \frac{z_2 - y_2}{z_2 - z_1} g(z_1) \geq g\left(\frac{z_2(y_2 - z_1) + z_1(z_2 - y_2)}{z_2 - z_1}\right) = g(y_2),
\]
\[
\frac{z_2 - y_2}{z_2 - z_1} g(z_2) + \frac{y_2 - z_1}{z_2 - z_1} g(z_1) \geq g\left(\frac{z_2(z_2 - y_2) + z_1(y_2 - z_1)}{z_2 - z_1}\right) = g(z_1 + z_2 - y_2) \geq g(y_1).
\]
Here the first two inequalities are Jensen while the last is by the first condition in (5) together with the monotonicity of \(g\). Adding these together yields \(g(z_2) + g(z_1) \geq g(y_2) + g(y_1)\), or \(E[g(Z)] \geq E[g(Y)]\), as desired. We need only check that the weights \(\frac{y_2 - z_1}{z_2 - z_1}\) and \(\frac{z_2 - y_2}{z_2 - z_1}\) are positive. The second weight is positive by the assumption \(z_2 \geq y_2\). The first weight could be negative if \(y_2 < z_1\), meaning that both \(y_1\) and \(y_2\) are smaller than both \(z_1\) and \(z_2\). But in this case, the inequality \(E[g(Y)] \leq E[g(Z)]\) trivially holds by monotonicity of \(g\). This completes the proof. \(\Box\)

We are now ready to prove Theorem 0.3.

**Proof of Theorem 0.3.** Without loss of generality we may take \(L = 1\). Our strategy will be to iteratively apply an inequality for a single \(\epsilon_i\) to replace all the \(\phi(t_i)\) with \(t_i\) one-by-one. The inequality for a single \(\epsilon_i\) is the following:

**Lemma 0.5.** For any 1-Lipschitz function \(\phi\) with \(\phi(0) = 0\), any collection \(T\) of ordered pairs \((a, b)\), and any convex increasing function \(g\), we have
\[
E_{\epsilon \sim \{-1,+1\}}[g(\sup_{(a,b)\in T} a + \epsilon \phi(b))] \leq E_{\epsilon \sim \{-1,+1\}}[g(\sup_{(a,b)\in T} a + \epsilon b)].
\]

To prove this, let \((a_+, b_+)\) attain the sup of \(a + \epsilon \phi(b)\) for \(\epsilon = +1\), and \((a_-, b_-)\) attain the sup for \(\epsilon = -1\).

We will check the conditions of Lemma 0.4 for
\[
y_1 = a_- - \phi(b_-),
\]
\[
y_2 = a_+ + \phi(b_+),
\]
\[
z_1 = \max(a_- - b_-, a_+ - b_+),
\]
\[
z_2 = \max(a_- + b_-, a_+ + b_+).
\]
(Note that \(z_1\) and \(z_2\) are lower-bounds on the right-hand-side sup for \(\epsilon = -1, +1\) respectively.)

First we need \(\max(y_1, y_2) \leq \max(z_1, z_2)\). But \(\max(z_1, z_2) = \max(a_+ + |b_-|, a_- + |b_+|) \geq \max(a_- - \phi(b_-), a_+ + \phi(b_+)) = \max(y_1, y_2)\). Here the inequality follows since \(\phi(b) \leq |b|\) since \(\phi\) is Lipschitz and \(\phi(0) = 0\).

Second we need \(\frac{y_1 + y_2}{2} \leq \frac{z_1 + z_2}{2}\). We have \(z_1 + z_2 \geq \max((a_- - b_-) + (a_+ + b_+), (a_- + b_-) + (a_+ - b_+)) = a_+ + a_- + |b_+ - b_-|\), so it suffices to show that \(\frac{a_+ + a_- + |b_+ - b_-|}{2} \geq \frac{a_+ + a_- + \phi(b_+) - \phi(b_-)}{2}\). This exactly reduces to \(\phi(b_+) - \phi(b_-) \leq |b_+ - b_-|\), which again follows since \(\phi\) is Lipschitz. This completes the proof of the lemma.

Now to prove the general proposition we observe that if \(g(x)\) is convex in \(x\), so is \(g(x + t)\) for any \(t\). We
then proceed by iteratively applying Lemma 0.5:

\[
\mathbb{E}_{\epsilon_1, n} [g(\sup_{t \in T} \sum_{i=1}^{n} \epsilon_i \phi(t_i))] = \mathbb{E}_{\epsilon_1, n-1} \left[ \mathbb{E}_{\epsilon_n} [g(\sup_{t \in T} \sum_{i=1}^{n-1} \epsilon_i \phi(t_i) + \epsilon_n \phi(t_n)) | \epsilon_1:n-1] \right]
\]

(13)

\[
\leq \mathbb{E}_{\epsilon_1, n-1} \left[ \mathbb{E}_{\epsilon_n} [g(\sup_{t \in T} \sum_{i=1}^{n-1} \epsilon_i \phi(t_i) + \epsilon_n t_n) | \epsilon_1:n-1] \right]
\]

(14)

\[
= \mathbb{E}_{\epsilon_1, n} [g(\sup_{t \in T} \sum_{i=1}^{n} \epsilon_i \phi(t_i) + \epsilon_n t_n)]
\]

(15)

\[
\vdots
\]

(16)

\[
\leq \mathbb{E}_{\epsilon_1, n} [g(\sup_{t \in T} \sum_{i=1}^{n} \epsilon_i t_i)]
\]

(17)

\[
\leq \mathbb{E}_{\epsilon_1, n} [g(\sup_{t \in T} \sum_{i=1}^{n} \epsilon_i t_i)],
\]

(18)

which completes the proof.

\[
\square
\]

Let us return now to bounding the truncated moments in Proposition 0.2.

**Proof of Proposition 0.2.** We start with a symmetrization argument. Let \( \mu_{\tilde{\psi}} = \mathbb{E}_{X \sim p^*}[\tilde{\psi}(\langle X - \mu, v \rangle) / \sigma] \), and note that \( \mu_{\tilde{\psi}} \leq \mu_{\psi} \leq 1 \). Now, by symmetrization we have

\[
\mathbb{E}_{X_1, \ldots, X_n \sim p^*} \left[ \frac{1}{n} \sum_{i=1}^{n} \tilde{\psi} \left( \frac{\langle X_i - \mu, v \rangle}{\sigma} \right) - \mu_{\tilde{\psi}} \right]^k
\]

(19)

\[
\leq \mathbb{E}_{X, X' \sim p^*, \epsilon} \left[ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \left( \tilde{\psi} \left( \frac{\langle X_i - \mu, v \rangle}{\sigma} \right) - \tilde{\psi} \left( \frac{\langle X'_i - \mu, v \rangle}{\sigma} \right) \right) \right]^k
\]

(20)

\[
\leq 2^k \mathbb{E}_{X \sim p^*} \left[ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \tilde{\psi} \left( \frac{\langle X_i - \mu, v \rangle}{\sigma} \right) \right]^k.
\]

(21)

Here the first inequality adds and subtracts the mean, the second applies symmetrization, while the third uses the fact that optimizing a single \( v \) for both \( X \) and \( X' \) is smaller than optimizing \( v \) separately for each (and that the expectations of the expressions with \( X \) and \( X' \) are equal to each other in that case).

We now apply Ledoux-Talagrand contraction. Invoking Theorem 0.3 with \( g(x) = |x|^k, \phi(x) = \tilde{\psi}(|x|) \) and \( t_i = \langle X_i - \mu, v \rangle / \sigma \), we obtain

\[
\mathbb{E}_{X \sim p^*} \left[ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \tilde{\psi} \left( \frac{\langle X_i - \mu, v \rangle}{\sigma} \right) \right]^k \leq (L/\sigma)^k \mathbb{E}_{X \sim p^*, \epsilon} \left[ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \langle X_i - \mu, v \rangle \right]^k
\]

(22)

\[
= (L/\sigma)^k \mathbb{E}_{X \sim p^*, \epsilon} \left[ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \langle X_i - \mu \rangle \right]^k.
\]

(23)

We are thus finally left to bound \( \mathbb{E}_{X \sim p^*} \left[ \sum_{i=1}^{n} \epsilon_i (X_i - \mu) \right]^k \). Here we will use Khintchine’s inequality, which says that

\[
A_k \|z\|_2 \leq \mathbb{E}_z [\|\sum_i \epsilon_i z_i \|^k]^{1/k} \leq B_k \|z\|_2,
\]

(24)

where \( A_k \) is \( \Theta(1) \) and \( B_k = \Theta(\sqrt{k}) \) for \( k \geq 1 \). Applying this in our case, we obtain

\[
\mathbb{E}_{X, \epsilon} \left[ \sum_{i=1}^{n} \epsilon_i (X_i - \mu) \right]^k \leq O(1)^k \mathbb{E}_{X, \epsilon, \epsilon'} \left[ \sum_{i=1}^{n} \epsilon_i (X_i - \mu, \epsilon') \right]^k.
\]

(25)
Next apply Rosenthal’s inequality (Eq. ??), which yields that

$$\mathbb{E}_{X, \epsilon} \sum_{i=1}^{n} \epsilon_i (X_i - \mu, \epsilon') | \epsilon' | \leq O(k)^{k} \mathbb{E}_{X, \epsilon} \sum_{i=1}^{n} | (X_i - \mu, \epsilon') | + O(\sqrt{k})^{k} \sum_{i=1}^{n} \mathbb{E}[(X_i - \mu, \epsilon')^2]^{k/2} \quad (26)$$

$$\leq O(k)^{k} \cdot n \sigma^k ||\epsilon'||^2 + O(\sqrt{k})^{k} \sigma^k ||\epsilon'||^2 \quad (27)$$

$$= O(\sigma k \sqrt{d})^{k} n + O(\sigma \sqrt{kd})^{k} n^{k/2}, \quad (28)$$

where the last step uses that $||\epsilon'||^2 = \sqrt{d}$ and the second-to-last step uses the bounded moments of $X$. As long as $n \gg k^{(k-2)}$ the latter term dominates and hence plugging back into we conclude that

$$\mathbb{E}_{X, \epsilon} \sum_{i=1}^{n} \epsilon_i (X_i - \mu) ||\epsilon||_2^{1/k} = O(\sigma \sqrt{dn}). \quad (29)$$

Thus bounds the symmetrized truncated moments in (22-23) by $O(L \sqrt{kd/n})$, and plugging back into (21) completes the proof. □

**Application: isotropic Gaussians.** Next take $\mathcal{G}_{\text{gauss}}$ to be the family of isotropic Gaussians $\mathcal{N}(\mu, I)$. We saw earlier that the modulus $m(\mathcal{G}_{\text{gauss}}, \epsilon)$ was $O(\epsilon)$ for the mean estimation loss $L(p, \theta) = ||\theta - \mu(p)||_2$. Thus projecting onto $\mathcal{G}_{\text{gauss}}$ yields error $O(\epsilon)$ for mean estimation in the limit of infinite samples, but doesn’t work for finite samples since the TV distance to $\mathcal{G}_{\text{gauss}}$ will always be 1.

Instead we will project onto the set $\mathcal{G}_{\text{cov}}(\sigma) = \{p | ||\mathbb{E}[(X - \mu)(X - \mu)^\top]|| \leq \sigma^2\}$, for $\sigma^2 = O(1 + d/n + \log(1/\delta)/n)$. We already saw in Lemma ?? that when $p^*$ is (sub-)Gaussian the empirical distribution $p_n^*$ lies within this set. But the modulus of $\mathcal{G}_{\text{cov}}$ only decays as $O(\sqrt{\epsilon})$, which is worse than the $O(\epsilon)$ dependence that we had in infinite samples! How can we resolve this issue?

We will let $\mathcal{G}_{\text{iso}}$ be the family of distributions whose covariance is not only bounded, but close to the identity, and where moreover this holds for all $(1-\epsilon)$-subsets:

$$\mathcal{G}_{\text{iso}}(\sigma_1, \sigma_2) \overset{\text{def}}{=} \{p | ||\mathbb{E}_{r}[X - \mu]||_2 \leq \sigma_1 \text{ and } ||\mathbb{E}_{r}[(X - \mu)(X - \mu)^\top - I]|| \leq (\sigma_2)^2, \text{ whenever } r \leq \frac{p}{1-\epsilon}\} \quad (30)$$

The following improvement on Lemma ?? implies that $p_n^* \in \mathcal{G}_{\text{iso}}(\sigma_1, \sigma_2)$ for $\sigma_1 = O(\epsilon \sqrt{\log(1/\epsilon)})$ and $\sigma_2 = O(\sqrt{\epsilon \log(1/\epsilon)})$. [Note: the lemma below is wrong as stated. To be fixed.]

**Lemma 0.6.** Suppose that $X_1, \ldots, X_n$ are drawn independently from a sub-Gaussian distribution with sub-Gaussian parameter $\sigma$, mean 0, and identity covariance. Then, with probability $1 - \delta$ we have

$$\left| \left| \frac{1}{|S|} \sum_{i \in S} X_i X_i^\top - I \right| \right| \leq O\left( \sigma^2 \cdot \left( \epsilon \log(1/\epsilon) + \frac{d + \log(1/\delta)}{n} \right) \right), \quad (31)$$

$$\left| \left| \frac{1}{|S|} \sum_{i \in S} X_i \right| \right| \leq O\left( \sigma \cdot \left( \epsilon \sqrt{\log(1/\epsilon)} + \sqrt{\frac{d + \log(1/\delta)}{n}} \right) \right) \quad (32)$$

for all subsets $S \subseteq \{1, \ldots, n\}$ with $|S| \geq (1 - \epsilon)n$. In particular, if $n \gg d/(\epsilon^2 \log(1/\epsilon))$ then $\delta \leq \exp(-c n \log(1/\epsilon))$ for some constant $c$.

We will return to the proof of Lemma 0.6 later. For now, note that this means that $p_n^* \in \mathcal{G}'$ for

$$\mathcal{G}' = \mathcal{G}_{\text{iso}}(O(\epsilon \sqrt{\log(1/\epsilon)}), O(\sqrt{\epsilon \log(1/\epsilon)})), \text{ at least for large enough } n. \text{ Furthermore, } \mathcal{G}' \subset \mathcal{M} \text{ for } \mathcal{M} = \mathcal{G}_{\text{cov}}(1 + O(\epsilon \log(1/\epsilon))).$$

Now we bound the generalized modulus of continuity:

**Lemma 0.7.** Suppose that $p \in \mathcal{G}_{\text{iso}}(\sigma_1, \sigma_2)$ and $q \in \mathcal{G}_{\text{cov}}(\sqrt{1 + \sigma_2^2})$, and furthermore $TV(p, q) \leq \epsilon$. Then $||\mu(p) - \mu(q)||_2 \leq O(\sigma_1 + \sigma_2 \sqrt{\epsilon} + \epsilon)$. 5
Proof. Take the midpoint distribution \( r = \frac{\min(p, q)}{1+\epsilon} \), and write \( q = (1-\epsilon)r + \epsilon q' \). We will bound \( ||\mu(r) - \mu(q)||_2 \) (note that \( ||\mu(r) - \mu(p)||_2 \) is already bounded since \( p \in G_{iso} \)). We have that

\[
\text{Cov}_q[X] = (1-\epsilon)(\text{Cov}_r[X] + (\mu_q - \mu_r)(\mu_q - \mu_r)^\top) + \epsilon(\mu_q - \mu_q')(\mu_q - \mu_q')^\top.
\]

A computation yields \( \mu_q - \mu_q' = \frac{(1-\epsilon)^2}{\epsilon}(\mu_q - \mu_r) \). Plugging this into (35) and simplifying, we obtain that

\[
\text{Cov}_q[X] \succeq (1-\epsilon)(\text{Cov}_r[X] + (1/\epsilon)(\mu_q - \mu_r)(\mu_q - \mu_r)^\top).
\]

Now since \( \text{Cov}_r[X] \succeq (1-\epsilon_2^2)I \), we have \( ||\text{Cov}_q[X]|| \geq (1-\epsilon)(1-\epsilon_2^2) + (1/\epsilon)||\mu_q - \mu_r||_2^2 \). But by assumption \( ||\text{Cov}_q[X]|| \leq 1+\epsilon_2^2 \). Combining these yields that \( ||\mu_q - \mu_q'||_2 \leq (2\epsilon^2 + \epsilon + \epsilon_2^2) \), and so \( ||\mu_r - \mu_q||_2 \leq \mathcal{O}(\epsilon_2 \sqrt{\epsilon}) \), which gives the desired result.

In conclusion, projecting onto \( G_{cov}(1 + \mathcal{O}(\epsilon \log(1/\epsilon))) \) under TV distance gives a robust mean estimator for isotropic Gaussians, which achieves error \( \mathcal{O}(\epsilon \sqrt{\log(1/\epsilon)}) \). This is slightly worse than the optimal \( \mathcal{O}(\epsilon) \) bound but improves over the naïve analysis that only gave \( \mathcal{O}(\sqrt{\epsilon}) \).

Another advantage of projecting onto \( G_{cov} \) is that, as we will see in Section ??, this projection can be done computationally efficiently.

Proof of Lemma 0.6. TBD