Quantization of Compressed Sensing Measurements

Özgür Yılmaz

Department of Mathematics
University of British Columbia

Jun 11, 2010
Collaborators

Joint work with:

- Sinan Güntürk (Courant)
- Mark Lammers (UNC Wilmington)
- Alex Powell (Vanderbilt)
- Rayan Saab (UBC)
Motivation: Digital Signal Processing

Inherently analog signals: Audio, images, seismic, etc.

Objective: Use digital technology to store and process analog signals – find efficient digital representations of analog signals.

How this is done - classical approach:

Signal $f$ (analog) $\rightarrow$ Sampling (I) $\rightarrow$ Quantization (II) $\rightarrow$ Compression (III)

A/D conversion: measurement & truncation

Source coding: truncation & compression (or other processing)
Motivation of CS: If the signal is sparse or compressible, can we combine sampling and compression stages to one compressed sensing stage?

What do we have so far? Sparse signals can be recovered from few non-adaptive, linear measurements. Major dimensionality reduction.

Missing link: Devise efficient quantization schemes for “compressed measurements”.

- Crucial if we want to compress.
- Little is known...
- This is the subject of this talk.
Compressed sensing

Notations.

- \( x \in \mathbb{R}^N \) is \textit{k-sparse} if \( x \) has at most \( k \) non-zero entries.
- \( \Sigma_k := \{ x \in \mathbb{R}^N : x \text{ is } k\text{-sparse} \} \)
- Measurement matrix: \( \Phi \), an \( m \times N \) real matrix.
- Measurements: \( y = \Phi x \) (\( \hat{y} = \Phi x + e \))
- Dimensional setting: \( k < m < N \).

Objective of CS. Suppose \( x \in \Sigma_k \) or can be well approximated from \( \Sigma_k \). Given the (noisy) measurements \( \hat{y} = \Phi x + e \),
- recover \( x \) exactly (approximately),
- in a computationally efficient manner.
The following “robust recovery” result is crucial for applications of CS (and our results in this talk).

**Theorem [Candès-Romberg-Tao], also [Donoho]**

Assume \(x \in \Sigma_k\) and \(\Phi\) satisfies RIP\((k, \delta)\) for sufficiently small \(\delta\). Let \(\hat{y} = \Phi x + e\) where \(\|e\|_2 \leq \epsilon\). Define the decoder (BPDN)

\[
\Delta_1^\epsilon(\hat{y}) := \text{arg min } \|z\|_1 \text{ subject to } \|\Phi z - \hat{y}\|_2 \leq \epsilon.
\]

Then

\[
x^\# = \Delta_1^\epsilon(\hat{y}) \implies \|x - x^\#\|_2 \leq C_0 \epsilon.
\]

More generally, for any \(x \in \mathbb{R}^N\), \(\|x - x^\#\|_2 \leq C_1 \epsilon + C_2 \frac{\sigma_k(x) \epsilon_1}{\sqrt{k}}\).

**Remarks.**

1. Gaussian random matrices \(\Phi\) with \(\Phi_{i,j} \sim \mathcal{N}(0, 1/m)\) satisfy RIP of desired order.

2. Potential sources for measurement error include quantization error.
Quantization of compressed sensing measurements

It is clear that compressed sensing is very effective for **dimension reduction**. However, one of the initial goals was to **compress**.

**Need efficient quantization strategies!**

**Setting.** Suppose $x \in \Sigma_k$, $\Phi$ a compressed sensing matrix. Let $y = \Phi x$ be the measurement of $x$.

**Problem.** Given a (discrete) quantization alphabet $\mathcal{A}$, e.g., $\mathcal{A} = d\mathbb{Z}$ (stick to this alphabet throughout this talk),

$$
x \in \mathbb{R}^N \xrightarrow{\Phi} y \in \mathbb{R}^m \xrightarrow{Q} q \in \mathcal{A}^m \xrightarrow{\Delta Q} x^\# \in \mathbb{R}^N
$$
Quantization of compressed sensing measurements

It is clear that compressed sensing is very effective for *dimension reduction*. However, one of the initial goals was to *compress*.

**Need efficient quantization strategies!**

**Setting.** Suppose \( x \in \Sigma_k \), \( \Phi \) a compressed sensing matrix. Let \( y = \Phi x \) be the measurement of \( x \).

**Problem.** Given a (discrete) quantization alphabet \( A \), e.g., \( A = d\mathbb{Z} \) (stick to this alphabet throughout this talk),

\[
\begin{align*}
x \in \mathbb{R}^N & \xrightarrow{\Phi} y \in \mathbb{R}^m & \xrightarrow{Q} q \in A^m & \xrightarrow{\Delta Q} x^\# \in \mathbb{R}^N
\end{align*}
\]
Quantization of compressed sensing measurements

It is clear that compressed sensing is very effective for dimension reduction. However, one of the initial goals was to compress.

Need efficient quantization strategies!

Setting. Suppose $x \in \Sigma_k$, $\Phi$ a compressed sensing matrix. Let $y = \Phi x$ be the measurement of $x$.

Problem. Given a (discrete) quantization alphabet $\mathcal{A}$, e.g., $\mathcal{A} = d\mathbb{Z}$ (stick to this alphabet throughout this talk),

$$
\begin{align*}
  x &\in \mathbb{R}^N &\Phi &\rightarrow & y &\in \mathbb{R}^m &\rightarrow & q &\in \mathcal{A}^m &\rightarrow & \Delta Q &\rightarrow & x^\# &\in \mathbb{R}^N \\
\end{align*}
$$

i.e., find a quantizer $Q : \mathbb{R}^m \leftrightarrow \mathcal{A}^m$ and a decoder $\Delta Q : \mathcal{A}^m \leftrightarrow \mathbb{R}^N$ such that

$\quad \|x - \Delta Q(q)\|$ is small whenever $x \in \Sigma_k$, and

$\quad \Delta Q$ is computationally tractable.
Quantization of CS measurements – PCM

The most intuitive quantizer. Round each measurement $y_j$ to the nearest element of $A = d\mathbb{Z}$ (Pulse Code Modulation or PCM)

$$Q_{PCM} : y \mapsto q_{PCM}.$$ 

Note that, setting $x_{#PCM} := \Delta^\epsilon_1(q_{PCM})$,

$$\|y - q_{PCM}\|_2 \leq \frac{d}{2} \sqrt{m} \text{ robust recovery} \Rightarrow \|x - x_{#PCM}\|_2 \lesssim d \sqrt{m}$$
Quantization of CS measurements – PCM

The most intuitive quantizer. Round each measurement $y_j$ to the nearest element of $\mathcal{A} = d\mathbb{Z}$ (Pulse Code Modulation or PCM)

$$Q_{\text{PCM}} : y \mapsto q_{\text{PCM}}.$$ 

Note that, setting $x^\#_{\text{PCM}} := \Delta_{1}(q_{\text{PCM}})$,

$$\|y - q_{\text{PCM}}\|_2 \leq \frac{d}{2} \sqrt{m} \quad \text{robust recovery} \quad \|x - x^\#_{\text{PCM}}\|_2 \lesssim d \sqrt{m}$$

This is counter-intuitive: more measurements $\implies$ more error?
Quantization of CS measurements – PCM

The most intuitive quantizer. Round each measurement $y_j$ to the nearest element of $A = d\mathbb{Z}$ (Pulse Code Modulation or PCM)

$$Q_{PCM} : y \mapsto q_{PCM}.$$ 

Note that, setting $x_{#PCM} := \Delta^\epsilon_1(q_{PCM}),$

$$\|y - q_{PCM}\|_2 \leq \frac{d}{2}\sqrt{m} \quad \text{robust recovery} \quad \|x - x_{#PCM}\|_2 \lesssim d\sqrt{m}$$

This is counter-intuitive: more measurements $\implies$ more error?

Not really! This is an artifact of our RIP-based choice of normalization for the measurement matrix $\Phi$, i.e., for $\Phi_{i,j} \sim \mathcal{N}(0, 1/m)$. Dynamic range of measurements vary with $m$, conflicts with fixed $A$. 
Quantization of CS measurements – PCM

**The most intuitive quantizer.** Round each measurement $y_j$ to the nearest element of $A = d\mathbb{Z}$ (Pulse Code Modulation or PCM)

$$Q_{PCM} : y \mapsto q_{PCM}.$$ 

Note that, setting $x_{PCM}^\# := \Delta^\epsilon_1(q_{PCM}),$

$$\|y - q_{PCM}\|_2 \leq \frac{d}{2} \sqrt{m} \quad \text{robust recovery} \quad \|x - x_{PCM}^\#\|_2 \lesssim d \sqrt{m}$$

**This is counter-intuitive:** more measurements $\implies$ more error?

Not really! This is an artifact of our RIP-based choice of normalization for the measurement matrix $\Phi$, i.e., for $\Phi_{i,j} \sim \mathcal{N}(0, 1/m)$. Dynamic range of measurements vary with $m$, conflicts with fixed $A$.

**Fixed** if we work with $\Phi$ such that $\Phi_{ij} \sim \mathcal{N}(0, 1)$. Then

$$\|y - q_{PCM}\|_2 \leq \frac{d}{2} \sqrt{m} \quad \text{robust recovery} \quad \|x - x_{PCM}^\#\|_2 \lesssim d$$
Still, the accuracy of reconstruction given by $\|x_{\text{PCM}}^{\#}\| \lesssim d$.

No improvement if we increase the number of measurements $m$!

Of course, this is just an upper bound, but...
Still, the accuracy of reconstruction given by $\|x_{\text{PCM}}^\#\| \lesssim d$.

No improvement if we increase the number of measurements $m$!

Of course, this is just an upper bound, but...

![Graph showing performance of various quantization/decoding schemes, $k = 10$]
Still, the accuracy of reconstruction given by $\|x^\#_{\text{PCM}}\| \lesssim d$.

No improvement if we increase the number of measurements $m$!

Of course, this is just an upper bound, but...

![Performance of various quantization/decoding schemes, $k = 10$](image)

Why should we expect the approximation to improve? If (once) the support of $x$ is known (recovered), then we have effectively oversampled $x$ ($m > k$ measurements for a $k$-dimensional signal)!

In traditional oversampled quantization theory, we can do significantly better!
Quantization of CS measurements – PCM

- Still, the accuracy of reconstruction given by $\|x^\#_{PCM}\| \lesssim d$.
- No improvement if we increase the number of measurements $m$!
- Of course, this is just an upper bound, but...

![Performance of various quantization/decoding schemes, $k = 10$](chart.png)

Why should we expect the approximation to improve?

If (once) the support of $x$ is known (recovered), then we have effectively oversampled $x$ ($m > k$ measurements for a $k$-dimensional signal)!

- In traditional oversampled quantization theory, we can do significantly better!
Still, the accuracy of reconstruction given by $\|x_{PCM}^\#\| \lesssim d$.

No improvement if we increase the number of measurements $m$!

Of course, this is just an upper bound, but...

Why should we expect the approximation to improve?
If (once) the support of $x$ is known (recovered), then we have effectively oversampled $x$ ($m > k$ measurements for a $k$-dimensional signal)!
Quantization of CS measurements – PCM

- Still, the accuracy of reconstruction given by $\|x_{PCM}^\#\| \lesssim d$.
- No improvement if we increase the number of measurements $m$!
- Of course, this is just an upper bound, but...

![Graph showing performance of various quantization/decoding schemes, $k = 10$](image)

**Why should we expect the approximation to improve?**
If (once) the support of $x$ is known (recovered), then we have effectively oversampled $x$ ($m > k$ measurements for a $k$-dimensional signal)!

- In traditional oversampled quantization theory, we can do **significantly better**!
Compressed sensing: Undersampled or oversampled?

Consider

\[
\begin{bmatrix}
* \\
* \\
* \\
* \\
* \\
\end{bmatrix}
= \begin{bmatrix}
\begin{array}{cccc}
3 & 4 & 6 \\
\end{array}
\end{bmatrix}
\begin{bmatrix}
\begin{array}{c}
0 \\
0 \\
0 \\
0 \\
0 \\
\end{array}
\end{bmatrix}
\]

If (once) the support \( T = \{3, 4, 6\} \) is known (recovered)

\[
\begin{bmatrix}
* \\
* \\
* \\
* \\
* \\
\end{bmatrix}
= \begin{bmatrix}
\begin{array}{c}
\begin{array}{ccc}
\text{green} & \text{blue} & \text{red}
\end{array}
\end{array}
\end{bmatrix}
\begin{bmatrix}
\begin{array}{c}
\begin{array}{c}
\text{green} \\
\text{blue} \\
\text{red}
\end{array}
\end{array}
\end{bmatrix}
\]

Observe:

- Rows of \( \Phi_T \) is a frame for \( \mathbb{R}^k \) with \( m > k \) vectors.
- Measurements \( y_j \) are associated frame coefficients.
- This is a redundant frame quantization problem.
Perspective and limitations of PCM

▶ For processing purposes, one should keep in mind that the CS measurements are a redundant encoding of some low-dimensional signal.

▶ If a quantization scheme is not effective for quantizing redundant frame expansions, it will not be effective in the CS setting.

▶ The following theorem gives a lower bound on the PCM-approximation error even if the support of sparse $x$ is known and even if the reconstruction is done optimally.

**Theorem (Goyal-Vetterli-Thao)**

Let $E$ be an $m \times k$ real matrix, and let $\Delta_{opt}$ be an optimal decoder. Then

$$\mathbb{E} \left\| x - \Delta_{opt} \left( Q_{PCM}(Ex) \right) \right\|_2^2 \gtrsim k \lambda^{-1} d$$

where the “oversampling rate” $\lambda := m/k$ and the expectation is wrt a probability measure on a bounded $K \subset \mathbb{R}^k$ that is, e.g., abs. continuous.
General frame quantization

Finite frames. A collection $E = \{e_j\}_1^m$ in $\mathbb{R}^k$ is a frame for $\mathbb{R}^k$ if

$$A\|x\|_2^2 \leq \sum_{j=1}^m |\langle x, e_j \rangle|^2 \leq B\|x\|_2^2, \quad \forall x \in \mathbb{R}^k.$$ 

Identify $E$ with the $m \times k$ matrix $E$ whose rows are $e_1^T, \ldots, e_m^T$. Then,

- $E$ is a frame for $\mathbb{R}^k$ if and only if $E$ is full rank.
- The frame bounds: $A = \sigma_{\min}^2(E)$ and $B = \sigma_{\max}^2(E)$.
- The entries of $y = Ex$ are the frame coefficients of $x \in \mathbb{R}^k$.
- The columns of any left inverse $F$ of $E$ is a dual frame of $E$. In this case, we can reconstruct $x$ from $y$ via $x = Fy$.
- The Moore-Penrose pseudo-inverse of $E$, given by

$$F_{\text{can}} = E^\dagger = (E^*E)^{-1}E^*$$

is the canonical dual of $E$. Then,
General frame quantization

**Problem.** Let $E$ be a frame for $\mathbb{R}^k$. Want to quantize the frame coefficients $y = Ex$ of $x \in \mathbb{R}^k$: As before, let $A = d\mathbb{Z}$, and let $Q$ be a quantizer, i.e.,

$$Q : y \in \mathbb{R}^m \mapsto q \in A^m.$$

Pick any left inverse (dual) $F$ of $E$ and set $\hat{x} = Fq$. Then

$$\text{approximation error} = x - \hat{x} = F(y - q).$$

**Two criteria for quantizer design**

- **“Noise shaping”** in this context: choose $q$ such that $y - q$ is close to $\text{Ker}(F)$.
- **Which dual $F$ of $E$?**
Fix the quantization method as $r$th-order $\Sigma\Delta$ quantization:

$$(\Delta^r u)_j = y_j - q_j.$$ 

Here $q_j$ can be chosen using the greedy rule which minimizes $u_j$ given $u_{j-1}, \ldots, u_{j-r}$ and $y_j$ if $\mathcal{A}$ is sufficiently large. In this case,

$$y - q_{\Sigma\Delta} = D^r u, \quad \text{with } \|u\|_\infty \leq d/2$$

where

$$D = \begin{bmatrix}
1 & 0 & 0 & 0 & \cdots & 0 \\
-1 & 1 & 0 & 0 & \cdots & 0 \\
0 & -1 & 1 & 0 & \cdots & 0 \\
0 & 0 & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & \cdots & -1 & 1 & 0 \\
0 & 0 & \cdots & 0 & -1 & 1
\end{bmatrix}_{m \times m}$$

Note also that $\|y - q_{\Sigma\Delta}\|_\infty \lesssim_r d$ (this will be important later).
From last slide: \( y - q_{\Sigma \Delta} = D^r u \) with \( \|u\|_\infty \lesssim d \).

Reconstruct with some dual \( F \) of \( E \): \( \hat{x}_{\Sigma \Delta} = Fq_{\Sigma \Delta} \). Then

\[
x - \hat{x}_{\Sigma \Delta} = F(y - q_{\Sigma \Delta}) = F D^r u.
\]
From last slide: \( y - q_{\Sigma \Delta} = D^r u \) with \( \|u\|_\infty \lesssim d \).

**Reconstruct** with some dual \( F \) of \( E \): \( \hat{x}_{\Sigma \Delta} = F q_{\Sigma \Delta} \). Then

\[
x - \hat{x}_{\Sigma \Delta} = F(y - q_{\Sigma \Delta}) = F D^r u.
\]

One way to bound the approximation error:

\[
\|x - \hat{x}_{\Sigma \Delta}\| \leq \|u\|_\infty \sum_j \| (FD^r)_j \|
\]
From last slide: \( y - q_{\Sigma \Delta} = D^r u \) with \( \|u\|_\infty \lesssim d \).

Reconstruct with some dual \( F \) of \( E \): \( \hat{x}_{\Sigma \Delta} = F q_{\Sigma \Delta} \). Then

\[
x - \hat{x}_{\Sigma \Delta} = F(y - q_{\Sigma \Delta}) = F D^r u.
\]

One way to bound the approximation error:

\[
\|x - \hat{x}_{\Sigma \Delta}\| \leq \|u\|_\infty \sum \|(FD')_j\| \leftarrow \text{frame variation}
\]

- Use \( \Sigma \Delta \) quantization if \( E \) admits a dual frame \( F \) whose columns vary smoothly. Accuracy: \( O(\lambda^{-r}) \) when \( r = 1, 2 \) where \( \lambda = m/k \). (Benedetto-Powell-Y)
From last slide: \( y - q_{\Sigma \Delta} = D^r u \) with \( \|u\|_\infty \lesssim d \).

Reconstruct with some dual \( F \) of \( E \): \( \hat{x}_{\Sigma \Delta} = F q_{\Sigma \Delta} \). Then

\[
x - \hat{x}_{\Sigma \Delta} = F(y - q_{\Sigma \Delta}) = F D^r u.
\]

One way to bound the approximation error:

\[
\|x - \hat{x}_{\Sigma \Delta}\| \leq \|u\|_\infty \sum \|(F D^r)_j\| \leftarrow \text{frame variation}
\]

- Use \( \Sigma \Delta \) quantization if \( E \) admits a dual frame \( F \) whose columns vary smoothly. Accuracy: \( O(\lambda^{-r}) \) when \( r = 1, 2 \) where \( \lambda = m/k \). (Benedetto-Powell-Y)

- Even \( \Sigma \Delta \) schemes of order 1 and 2 are significantly superior to PCM. (BPY, later by Bodmann-Paulsen, Boufounos-Oppenheim,...).
**ΣΔ quantization for finite frames – approximation error**

From last slide: \( y - q_{ΣΔ} = D^r u \) with \( \|u\|_∞ \lesssim d \).

**Reconstruct** with some dual \( F \) of \( E \): \( \hat{x}_{ΣΔ} = Fq_{ΣΔ} \). Then

\[
x - \hat{x}_{ΣΔ} = F(y - q_{ΣΔ}) = FD^r u.
\]

One way to bound the approximation error:

\[
\|x - \hat{x}_{ΣΔ}\| \leq \|u\|_∞ \sum \|(FD^r)_j\| \leftarrow \text{frame variation}
\]

- Use ΣΔ quantization if \( E \) admits a dual frame \( F \) whose columns vary smoothly. Accuracy: \( O(\lambda^{-r}) \) when \( r = 1, 2 \) where \( \lambda = m/k \). (Benedetto-Powell-Y)

- Even ΣΔ schemes of order 1 and 2 are significantly superior to PCM. (BPY, later by Bodmann-Paulsen, Boufounos-Oppenheim, ...).

- In these early results, the focus is on canonical-dual reconstructions.
Problem 1. Extension to higher-order schemes is non-trivial! (Negative result for, e.g., harmonic frames together with their canonical duals for schemes of order 3 or higher by Lammers-Powell-Y.)
**Problem 1.** Extension to higher-order schemes is non-trivial! (Negative result for, e.g., harmonic frames together with their canonical duals for schemes of order 3 or higher by Lammers-Powell-Y.)

**Remedy.** Construct suitable alternative duals!

**How?** If we work with the $\ell_2$ norm:

$$\|x - \hat{x}_{\Sigma\Delta}\|_2 \leq \|FD^r\|_{op}\|u\|_2.$$ 

**Seek** a dual $F$ that minimizes $\|FD^r\|_{op}$.

**Theorem [BLPY].** If $E$ is sufficiently smooth (e.g., sampled from a piecewise-$C^1$ frame path), an $r$th-order $\Sigma\Delta$ scheme produces an accuracy of $O(\lambda^r)$ if the reconstruction is done using $F_{sob}^r$, given explicitly by

$$F_{sob}^r = (D - rE)^\dagger(D - rE).$$
**Problem 1.** Extension to higher-order schemes is non-trivial! (Negative result for, e.g., harmonic frames together with their canonical duals for schemes of order 3 or higher by Lammers-Powell-Y.)

**Remedy.** Construct suitable alternative duals!

**How?** If we work with the $\ell_2$ norm:

$$\|x - \hat{x}_{\Sigma \Delta}\|_2 \leq \|FD^r\|_{\text{op}}\|u\|_2.$$

**Seek** a dual $F$ that minimizes $\|FD^r\|_{\text{op}}$.

**Solution** is the $r$th-order Sobolev dual of $E$, introduced by Blum-Lammers-Powell-Y, given explicitly by

$$F_{\text{sob}, r} = (D^{-r} E)^\dagger D^{-r}.$$
Problem 1. Extension to higher-order schemes is non-trivial! (Negative result for, e.g., harmonic frames together with their canonical duals for schemes of order 3 or higher by Lammers-Powell-Y.)

Remedy. Construct suitable alternative duals!

How? If we work with the $\ell_2$ norm:

$$\| x - \hat{x}_{\Sigma\Delta} \|_2 \leq \| F D^r \|_{op} \| u \|_2.$$ 

Seek a dual $F$ that minimizes $\| F D^r \|_{op}$.

Solution is the $r$th-order Sobolev dual of $E$, introduced by Blum-Lammers-Powell-Y, given explicitly by

$$F_{sob,r} = (D^{-r} E)^\dagger D^{-r}.$$ 

Theorem [BLPY].

If $E$ is sufficiently smooth (e.g., sampled from a piecewise-$C^1$ frame path), an $r$th-order $\Sigma\Delta$ scheme produces an accuracy of $O(\lambda^{-r})$ if the reconstruction is done using $F_{sob,r}$.
Problem 2. What if the original frame $E$ is not smooth, like our $\Phi_T$ which is a “Gaussian random frame”. 

This is our main problem. We will show that Sobolev duals can still be used for $\Sigma\Delta$ quantization of Gaussian random frames.
Problem 2. What if the original frame $E$ is not smooth, like our $\Phi_T$ which is a “Gaussian random frame”.

This is our main problem. We will show that Sobolev duals can still be used for $\Sigma\Delta$ quantization of Gaussian random frames.
Problem 2. What if the original frame $E$ is not smooth, like our $\Phi_T$ which is a “Gaussian random frame”.

This is our main problem. We will show that Sobolev duals can still be used for $\Sigma\Delta$ quantization of Gaussian random frames.
Problem 2. What if the original frame $E$ is not smooth, like our $\Phi_T$ which is a “Gaussian random frame”.

**This is our main problem.** We will show that Sobolev duals can still be used for $\Sigma\Delta$ quantization of Gaussian random frames.
**Σ∆ quantization for random frames – Sobolev duals**

**Problem 2.** What if the original frame $E$ is not smooth, like our $\Phi_T$ which is a “Gaussian random frame”.

**This is our main problem.** We will show that Sobolev duals can still be used for Σ∆ quantization of Gaussian random frames.

![Graphical representation of Gaussian frame and its Sobolev dual](image.png)

- Gaussian frame in $\mathbb{R}^2$
- Its Sobolev dual of order $r = 1$
A generic error bound.

Let $E$ be an $m \times k$ full-rank matrix. If $y = Ex$ is quantized via a (stable) $r$th-order $\Sigma\Delta$ scheme, and $\hat{x}_{\Sigma\Delta} = F_{\text{sob,}r}q$, then

$$
\|x - x_{\Sigma\Delta}\|_2 \lesssim_r \frac{d\sqrt{m}}{\sigma_{\min}(D - rE)}.
$$

In other words, we need to control $\sigma_{\min}(D - rE)$ as a function of the oversampling rate $\lambda = m/k$. 

A generic error bound.

Let $E$ be an $m \times k$ full-rank matrix. If $y = Ex$ is quantized via a (stable) $r$th-order $\Sigma\Delta$ scheme, and $\hat{x}_{\Sigma\Delta} = F_{\text{sob},rq}$, then

$$\|x - x_{\Sigma\Delta}\|_2 \lesssim_r d\sqrt{m} \frac{d\sqrt{m}}{\sigma_{\min}(D-rE)}.$$ 

In other words, we need to control $\sigma_{\min}(D-rE)$ as a function of the oversampling rate $\lambda = m/k$.

Rest of the talk: Focus on the case when $E$ is an $m \times k$ Gaussian random matrix, i.e., $E_{ij} \sim \mathcal{N}(0,1)$.

How does $\sigma_{\min}(D-rE)$ behave when $E$ is Gaussian random matrix depending on $m$, $k$, and $r$?
Least (non-zero) singular value of $D^{-r}E$.

Some important facts:

- For an $m \times k$ Gaussian $E$ (in fact, also for sub-Gaussian), it is known [Rudelson-Vershynin] that, with high probability,

  $$\sigma_{\text{min}}(E) \approx \sqrt{m} - \sqrt{k} - 1.$$
Least (non-zero) singular value of $D^{-r}E$.

Some important facts:

▶ For an $m \times k$ Gaussian $E$ (in fact, also for sub-Gaussian), it is known [Rudelson-Vershynin] that, with high probability,

$$\sigma_{\text{min}}(E) \approx \sqrt{m} - \sqrt{k} - 1.$$

▶ Singular values of $D$: explicitly known (related to DCT-VIII, e.g., Strang).
Least (non-zero) singular value of $D^{-r}E$.

Some important facts:

- For an $m \times k$ Gaussian $E$ (in fact, also for sub-Gaussian), it is known [Rudelson-Vershynin] that, with high probability,

$$\sigma_{\min}(E) \approx \sqrt{m} - \sqrt{k} - 1.$$ 

- Singular values of $D$: explicitly known (related to DCT-VIII, e.g., Strang).

- Eigenvalues of $D^* r D^r$ can be estimated from the eigenvalues of $(D^* D)^r$ using Weyl’s inequality. (Güntürk-Lammers-Powell-Saab-Y — GLPSY from now on).

- In particular,

$$\sigma_{\min}(D^{-r}) \approx r 1.$$
Try the product bound:
\[ \sigma_{\min}(D^{-r}E) \geq \sigma_{\min}(D^{-r})\sigma_{\min}(E). \]

Using the observations above, this gives:
\[ \|x - \hat{x}_{\Sigma\Delta}\|_2 \lesssim_r d, \]

This is not different from what we had from PCM-quantized measurements using \( \ell^1 \)-minimization decoding. Can’t we do better?
Try the product bound:

$$\sigma_{\text{min}}(D^{-r}E) \geq \sigma_{\text{min}}(D^{-r})\sigma_{\text{min}}(E).$$

Using the observations above, this gives:

$$\|x - \hat{x}_{\Sigma\Delta}\|_2 \lesssim_r d,$$

This is not different from what we had from PCM-quantized measurements using $\ell^1$-minimization decoding. Can’t we do better?

**Theorem I [GLPSY]**

Let $E_{ij} \sim \mathcal{N}(0,1)$. For any $\alpha \in (0,1)$, if $\lambda := \frac{m}{k} \gtrsim_{\alpha,r} (\log m)^{1/(1-\alpha)}$, then

$$\sigma_{\text{min}}(D^{-r}E) \gtrsim_r \lambda^{\alpha(r-\frac{1}{2})}\sqrt{m}$$

with probability at least $1 - \exp(-c'm\lambda^{-\alpha})$. Hence,

$$\|x - \hat{x}_{\Sigma\Delta}\|_2 \lesssim_r \frac{d}{\lambda^{\alpha(r-\frac{1}{2})}}.$$
Singular values of $\sigma_{\min}(D^{-r}E) \sim$ numerical experiment

$k = 50, 1 \leq \lambda \leq 25, E_{ij} \sim \mathcal{N}(0, 1), \frac{\sqrt{m}}{\sigma_{\min}(D^{-r}E)} \lesssim r \lambda^{-\alpha(r-\frac{1}{2})}$.
Main ingredients of the proof:

▷ Weyl’s inequality for the singular value estimates, in particular to estimate the singular values of $D^{-r}$.

▷ Unitary invariance of the i.i.d. Gaussian measure. Reduces the problem to estimating $\sigma_{\min}(\Sigma E)$ where $\Sigma$ is diagonal with $\Sigma_{ii}$ are estimated as described above.

▷ Concentration of measure for $\Sigma E$: estimate

$$\mathbb{P}\{\gamma \|x\|_2 \leq \|\Sigma E x\|_2 \leq \theta \|x\|_2\}.$$ 

▷ Pass to the singular values of $\Sigma E$ by using a standard net argument.
Implications for compressed sensing quantization

**Goal:** Use $\Sigma\Delta$ quantization for compressed sensing measurements. More precisely, with $k < m < N$, let:

- $\Phi$: an $m \times N$ Gaussian CS measurement matrix.
- $x \in \Sigma_k$: a $k$-sparse vector in $\mathbb{R}^N$, supported on $T$.
- $y = \Phi x$: CS measurement of $x$
- $q_{\Sigma\Delta} = \hat{y}$: $\Sigma\Delta$-quantized ($r$th-order) measurements.

**Problem 1.** Can we estimate the support $T$ of $x$ from $q_{\Sigma\Delta}$?

- If yes, then recall: $y = \Phi x = \Phi^T x^T$, where $\Phi^T$ is $m \times k$ Gaussian random matrix.
- Thus Theorem I applies for each $x$ with high probability.

**Problem 2.** Can we obtain a uniform guarantee, i.e., have Theorem I hold for each submatrix $\Phi^T$ (with high probability)?
Implications for compressed sensing quantization

**Goal:** Use $\Sigma\Delta$ quantization for compressed sensing measurements. More precisely, with $k < m < N$, let:

- $\Phi$: an $m \times N$ Gaussian CS measurement matrix.
- $x \in \Sigma_k$: a $k$-sparse vector in $\mathbb{R}^N$, supported on $T$.
- $y = \Phi x$: CS measurement of $x$
- $q_{\Sigma\Delta} = \hat{y}$: $\Sigma\Delta$-quantized ($r$th-order) measurements.

**Problem 1.** Can we estimate the support $T$ of $x$ from $q_{\Sigma\Delta}$?

- If yes, then recall: $y = \Phi x = \Phi_T x_T$, where $\Phi_T$ is $m \times k$ Gaussian random matrix.
- Thus Theorem I applies for each $x$ with high probability.
**Goal:** Use $\Sigma \Delta$ quantization for compressed sensing measurements. More precisely, with $k < m < N$, let:

- $\Phi$: an $m \times N$ Gaussian CS measurement matrix.
- $x \in \Sigma_k$: a $k$-sparse vector in $\mathbb{R}^N$, supported on $T$.
- $y = \Phi x$: CS measurement of $x$
- $q_{\Sigma \Delta} = \hat{y}$: $\Sigma \Delta$-quantized ($r$th-order) measurements.

**Problem 1.** Can we estimate the support $T$ of $x$ from $q_{\Sigma \Delta}$?

- If yes, then recall: $y = \Phi x = \Phi_T x_T$, where $\Phi_T$ is $m \times k$ Gaussian random matrix.

- Thus Theorem I applies for each $x$ with high probability.

**Problem 2.** Can we obtain a uniform guarantee, i.e., have Theorem I hold for each submatrix $\Phi_T$ (with high probability)?
Answer 1. Yes! This follows from the robust recovery result for the $\ell_1$ decoder.

Proposition.

If $|x_j| \gtrsim_r d$ for all $j \in T$, then the largest $k$ entries of $x^\# = \Delta_1^\epsilon(q\Sigma\Delta)$ are supported on $T$. 
**Answer 1.** Yes! This follows from the robust recovery result for the $\ell_1$ decoder.

**Proposition.**

If $|x_j| \gtrsim_r d$ for all $j \in T$, then the largest $k$ entries of $x^\# = \Delta_1^c (q_{\Sigma \Delta})$ are supported on $T$.

**Answer 2.** Yes! This follows using a union bound.

**Theorem II [GLPSY]**

Let $\Phi$ be a Gaussian random matrix. If $\lambda \gtrsim_{\alpha, r} (\log N)^{1/(1-\alpha)}$, Theorem I holds for $\Phi_T$ for all $T$ with $|T| \leq k$ w.h.p. on the draw of $\Phi$. 
In light of all these results, we propose ΣΔ quantization and a two-stage recovery procedure for compressed sensing:

1. **Coarse recovery:** \( \ell_1 \)-minimization (or any other robust recovery procedure) applied to \( q_{\Sigma\Delta} y \) yields an initial, "coarse" approximation \( x_\#(q_{\Sigma\Delta}) \) of \( x \), and in particular, the exact (or otherwise approximate) support \( T \) of \( x \).

2. **Fine recovery:** Sobolev dual of the frame \( \Phi_T \) applied to \( q_{\Sigma\Delta} y \) yields a finer, final approximation \( \hat{x}_{\Sigma\Delta} \) of \( x \).
In light of all these results, we propose ΣΔ quantization and a two-stage recovery procedure for compressed sensing:

1. **Coarse recovery**: $\ell_1$-minimization (or any other robust recovery procedure) applied to $q_{\Sigma\Delta}$ yields an initial, “coarse” approximation $x^\#(q_{\Sigma\Delta})$ of $x$, and in particular, the exact (or otherwise approximate) support $T$ of $x$. 
ΣΔ quantization for compressed sensing

In light of all these results, we propose ΣΔ quantization and a two-stage recovery procedure for compressed sensing:

1. **Coarse recovery**: $\ell_1$-minimization (or any other robust recovery procedure) applied to $q_{\Sigma\Delta}$ yields an initial, "coarse" approximation $x^\#(q_{\Sigma\Delta})$ of $x$, and in particular, the exact (or otherwise approximate) support $T$ of $x$.

2. **Fine recovery**: Sobolev dual of the frame $\Phi_T$ applied to $q_{\Sigma\Delta}$ yields a finer, final approximation $\hat{x}_{\Sigma\Delta}$ of $x$. 
ΣΔ quantization for compressed sensing

In light of all these results, we propose ΣΔ quantization and a two-stage recovery procedure for compressed sensing:

1. **Coarse recovery**: $\ell_1$-minimization (or any other robust recovery procedure) applied to $q_{\Sigma\Delta}$ yields an initial, “coarse” approximation $x^\#(q_{\Sigma\Delta})$ of $x$, and in particular, the exact (or otherwise approximate) support $T$ of $x$.

2. **Fine recovery**: Sobolev dual of the frame $\Phi_T$ applied to $q_{\Sigma\Delta}$ yields a finer, final approximation $\hat{x}_{\Sigma\Delta}$ of $x$.

Combining all these:
Σ∆ quantization for compressed sensing

In light of all these results, we propose Σ∆ quantization and a two-stage recovery procedure for compressed sensing:

1. **Coarse recovery:** $\ell_1$-minimization (or any other robust recovery procedure) applied to $q_{\Sigma\Delta}$ yields an initial, “coarse” approximation $x^\#(q_{\Sigma\Delta})$ of $x$, and in particular, the exact (or otherwise approximate) support $T$ of $x$.

2. **Fine recovery:** Sobolev dual of the frame $\Phi_T$ applied to $q_{\Sigma\Delta}$ yields a finer, final approximation $\hat{x}_{\Sigma\Delta}$ of $x$.

Combining all these:

**Theorem III [GLPSY]**

With high probability on the initial draw of $\Phi$ and uniformly for all $k$-sparse $x$ that satisfy the above size condition, we have

$$\|x - \hat{x}_{\Sigma\Delta}\|_2 \lesssim r \frac{d}{(m/k)^\alpha(r^{-1/2})}$$

if $m \gtrsim_{\alpha,r} k(\log N)^{1/(1-\alpha)}$. 
Decoder Performance on Sparse Signals

The graph shows the performance of various quantization/decoding schemes for different values of $\lambda$. The y-axis represents the mean $l_2$-norm of the error, while the x-axis represents $\lambda$. The graph compares different schemes, including PCM $\rightarrow l_1$, PCM $\rightarrow l_1 \rightarrow F_{can}$, $\Sigma\Delta$ with $r=1, 2, 3$, and $c \lambda^{-r} k^{1/2}$ with $r=0.5, 1, 2$. The error norm decreases as $\lambda$ increases, indicating improved performance of the decoding schemes.
Rate-distortion issues

Implications regarding the dependence of the approximation error on the bit-budget?

Need to use finite alphabet quantizers.

- Signals of interest in $K := \{ x \in \Sigma^N_k : A \leq |x_j| \leq \rho, \ \forall j \in T \}$, with $A \ll \rho$.

- We use a $B_r$-bit uniform quantizer with the largest allowable step-size, say $\delta_r$, for our support recovery result to hold.

- Above, choose $B_r = B_r(\rho)$ so that the associated $\Sigma \Delta$ quantizer does not overload

The approximation error (distortion) $D_{\Sigma \Delta}$ after the fine recovery stage via Sobolev duals:

$$D_{\Sigma \Delta} \lesssim_r \lambda^{-\alpha(r-1/2)} \delta_r \approx \frac{\lambda^{-\alpha(r-1/2)} A}{2^{r+1/2}}.$$
Rate-distortion issues

Implications regarding the dependence of the approximation error on the bit-budget?

Need to use finite alphabet quantizers.

- Signals of interest in $K := \{x \in \Sigma_k^N : A \leq |x_j| \leq \rho, \ \forall j \in T\}$, with $A \ll \rho$.
- We use a $B_r$-bit uniform quantizer with the largest allowable step-size, say $\delta_r$, for our support recovery result to hold.
- Above, choose $B_r = B_r(\rho)$ so that the associated $\Sigma\Delta$ quantizer does not overload.

The approximation error (distortion) $D_{\Sigma\Delta}$ after the fine recovery stage via Sobolev duals:

$$D_{\Sigma\Delta} \lesssim_r \lambda^{-\alpha(r-1/2)} \delta_r \approx \frac{\lambda^{-\alpha(r-1/2)}A}{2^{r+1/2}}.$$ 

How about PCM? Same step size $\delta_r$ along with $\ell_1$ decoder requires roughly the same number of bits $B_r$, but provides the distortion bound

$$D_{\text{PCM}} \lesssim \delta_r \approx \frac{A}{2^{r+1/2}}.$$
Rate-distortion issues

Implications regarding the dependence of the approximation error on the bit-budget?

Need to use finite alphabet quantizers.

- Signals of interest in \( K := \{ x \in \Sigma_k^N : A \leq |x_j| \leq \rho, \forall j \in T \} \), with \( A \ll \rho \).

- We use a \( B_r \)-bit uniform quantizer with the largest allowable step-size, say \( \delta_r \), for our support recovery result to hold.

- Above, choose \( B_r = B_r(\rho) \) so that the associated \( \Sigma \Delta \) quantizer does not overload

The approximation error (distortion) \( D_{\Sigma \Delta} \) after the fine recovery stage via Sobolev duals:

\[
D_{\Sigma \Delta} \lesssim_r \lambda^{-\alpha(r-1/2)} \delta_r \approx \frac{\lambda^{-\alpha(r-1/2)} A}{2^{r+1/2}}.
\]

How about PCM? Same step size \( \delta_r \) along with \( \ell_1 \) decoder requires roughly the same number of bits \( B_r \), but provides the distortion bound

\[
D_{\text{PCM}} \lesssim \delta_r \approx \frac{A}{2^{r+1/2}}.
\]
Pros

1. More accurate than any known quantization scheme in this setting (even when sophisticated recovery algorithms are employed).
2. Modular: If the fine recovery stage is not available or impractical, then the standard (coarse) recovery procedure is applicable as is.
3. Progressive: If new measurements arrive (in any given order), noise shaping can be continued on these measurements as long as the state of the system (real values for an $r$th order scheme) has been stored.
4. Universal: It uses no information about the measurement matrix or the signal.

(Potential) Cons

More work for reconstruction. Out-of-band noise sensitivity to be sorted out as well as extension to compressible signals (work in progress).
ΣΔ Quantization for Compressed Sensing

Pros

1. More accurate than any known quantization scheme in this setting (even when sophisticated recovery algorithms are employed).
Σ∆ Quantization for Compressed Sensing

Pros

1. **More accurate** than any known quantization scheme in this setting (even when sophisticated recovery algorithms are employed).

2. **Modular**: If the fine recovery stage is not available or impractical, then the standard (coarse) recovery procedure is applicable as is.

(Potential) Cons

More work for reconstruction. Out-of-band noise sensitivity to be sorted out as well as extension to compressible signals (work in progress).
Pros

1. **More accurate** than any known quantization scheme in this setting (even when sophisticated recovery algorithms are employed).

2. **Modular:** If the fine recovery stage is not available or impractical, then the standard (coarse) recovery procedure is applicable as is.

3. **Progressive:** If new measurements arrive (in any given order), noise shaping can be continued on these measurements as long as the state of the system ($r$ real values for an $r$th order scheme) has been stored.

(Potential) Cons

More work for reconstruction. Out-of-band noise sensitivity to be sorted out as well as extension to compressible signals (work in progress).
Pros

1. **More accurate** than any known quantization scheme in this setting (even when sophisticated recovery algorithms are employed).

2. **Modular**: If the fine recovery stage is not available or impractical, then the standard (coarse) recovery procedure is applicable as is.

3. **Progressive**: If new measurements arrive (in any given order), noise shaping can be continued on these measurements as long as the state of the system ($r$ real values for an $r$th order scheme) has been stored.

4. **Universal**: It uses no information about the measurement matrix or the signal.
**Pros**

1. **More accurate** than any known quantization scheme in this setting (even when sophisticated recovery algorithms are employed).

2. **Modular:** If the fine recovery stage is not available or impractical, then the standard (coarse) recovery procedure is applicable as is.

3. **Progressive:** If new measurements arrive (in any given order), noise shaping can be continued on these measurements as long as the state of the system \( r \) real values for an \( r \)th order scheme) has been stored.

4. **Universal:** It uses no information about the measurement matrix or the signal.

**(Potential) Cons**

More work for reconstruction. Out-of-band noise sensitivity to be sorted out as well as extension to compressible signals (work in progress).