Quantized Iterative Hard Thresholding: Bridging 1-bit and High Resolution Quantized Compressed Sensing

Laurent Jacques, Kévin Degraux, Christophe De Vleeschouwer

Louvain University (UCL), Louvain-la-Neuve, Belgium
Compressive Sampling and Quantization

Compressed sensing theory says:

“Linearly sample a signal at a rate function of its intrinsic dimensionality”

Information theory and sensor designer say:

“Okay, but I need to quantize/digitize my measurements!”
(e.g., in ADC)
Focus on scalar quantization

Turning measurements into bits → scalar quantization

Important conventions:

- Definition of $\Phi$ independent of $M$ (e.g., $\Phi_{ij} \sim_{\text{iid}} \mathcal{N}(0, 1)$) → preserves measurement dynamic!
- $B$ bits per measurement
- Total bit budget: $R = BM$
- No further encoding (e.g., entropic)

$$q_i = Q[(\Phi x)_i] = Q[\langle \phi_i, x \rangle] \in \Omega \subseteq \mathbb{R}$$

$$q = Q[\Phi x] \in \Omega = \Omega^M,$$

$$\Omega = \{q_i \in \mathbb{R} : 1 \leq i \leq 2^B\}, \quad \text{(levels)}$$

$$T = \{\tau_i \in \mathbb{R} : 1 \leq i \leq 2^B + 1, \tau_i \leq \tau_{i+1}\} \quad \text{(thresholds)}$$
1. First extreme:
CS and high-resolution scalar quantization
Former solutions

- Quantization is like a noise

\[ q = Q[\Phi x] = \Phi x + n \]
Former solutions

- Quantization is like a noise

\[ q = Q[\Phi x] = \Phi x + n \]

- CS is robust, *e.g.*, with ...
  - **Basis Pursuit DeNoi**se (BP**D**)N):
    \[ \hat{x} = \arg\min_{u \in \mathbb{R}^N} \| u \|_1 \text{ s.t. } \| \Phi u - q \| \leq \epsilon \]
  - **Iterative Hard Thresholding** (IHT):
    \[ \hat{x} = \arg\min_{u \in \mathbb{R}^N} \| \Phi u - y \|_2^2 \text{ s.t. } \| u \|_0 \leq K \]
    \[ \approx \text{ iterating: } x^{(n+1)} = \mathcal{H}_K(x^{(n)} + \mu \Phi^*(y - \Phi x^{(n)})) \]
Former solutions

- Quantization is like a noise
  \[ q = Q[\Phi x] = \Phi x + n \]

- CS is robust (e.g., with BPDN or IHT)
  If \( \|n\| \leq \epsilon \) and \( \frac{1}{\sqrt{M}} \Phi \) is RIP(\( \delta, aK \)) with \( \delta \leq \delta_0 \), then
  \[ \|\hat{x} - x\| \leq C \frac{\epsilon}{\sqrt{M}} + D e(K), \]
  for some \( C, D > 0 \) and \( e(K) = \text{deviation to } K\text{-sparsity.} \)

\[ e(K) = \|x - x_K\|_1/\sqrt{K} \text{ or } \|x - x_K\|_2 + \|x - x_K\|_1/\sqrt{K} \]

[Candès, 08] (\( a = 2, \delta_0 < \sqrt{2} - 1 \))
[Blumensath, Davies, 09] (\( a = 3, \delta_0 < 1/32 \))
**Former solutions**

- Quantization is like a noise

\[ q = Q[\Phi x] = \Phi x + n \]

- CS is robust (e.g., with BPDN or IHT)

If \( \|n\| \leq \epsilon \) and \( \frac{1}{\sqrt{M}} \Phi \) is RIP(\( \delta, aK \)) with \( \delta \leq \delta_0 \), then

\[ \|\hat{x} - x\| \leq C \frac{\epsilon}{\sqrt{M}} + D e(K), \]

for some \( C, D > 0 \) and \( e(K) = \text{deviation to } K\text{-sparsity.} \)

- **Question**: finding a \( \epsilon \) for QCS? High-Resolution Bounds!

\( e.g., \frac{\epsilon}{\sqrt{M}} = O(\alpha) \) (unif. \( Q \) bin width \( \alpha \)) or \( O(2^{-B}) \) for \( B\)-bit \( Q \)
2. Second extreme: CS and 1-bit scalar quantization
1-bit Compressed Sensing

\[ q = \text{sign} \Phi x \]

Oversampling in \( M \)

\( M \)-bits!

But, which information inside \( q \)?

[Boffounos, Baraniuk, 08]
1-bit Compressed Sensing

\[ \Phi \times q = \text{sign} \]

Oversampling in \( M \) \( \Rightarrow \) \( M \)-bits!

But, which information inside \( q \)?

[Bofounos, Baraniuk, 08]

bits matter!
1-bit Compressed Sensing

\[
q = \text{sign} \left[ \begin{bmatrix} \Phi \\ x \end{bmatrix} \right]
\]

**Warning 1:** signal amplitude is lost!

⇒ Amplitude is arbitrarily fixed

Examples: \( \| x \| = 1 \) or \( \| \Phi x \|_1 = 1 \)

**Warning 2:** \( \exists \) forbidden sensing!

(e.g. Bernoulli + canonical sparsity)

[Plan, Vershynin, 11]
Why does it work?

\( \mathbf{x} \) on \( S^2 \)

\( M \) vectors:
\[
\{ \varphi_i : 1 \leq i \leq M \}
\]

\text{iid Gaussian}

1-bit Measurements
\[
\begin{align*}
\langle \varphi_1, \mathbf{x} \rangle &> 0 \\
\langle \varphi_2, \mathbf{x} \rangle &> 0 \\
\langle \varphi_3, \mathbf{x} \rangle &\leq 0 \\
\langle \varphi_4, \mathbf{x} \rangle &> 0 \\
\langle \varphi_5, \mathbf{x} \rangle &> 0 \\
\vdots &
\end{align*}
\]

Smaller and smaller when \( M \) increases
\[
\{ \mathbf{u} : \text{sign}(\Phi \mathbf{u}) = \text{sign}(\Phi \mathbf{x}) \}
\]

Lower bound on this width?
1-bit CS bounds

Let $A(\cdot) := \text{sign}(\Phi \cdot)$ with $\Phi \sim \mathcal{N}^{M \times N}(0, 1)$.

If $M = O(\epsilon^{-1} K \log N)$, then, w.h.p,
for any two unit $K$-sparse vectors $x$ and $s$,

if only $b$ bits are different between $A(x)$ and $A(s)$

$\Rightarrow \|x - s\| \leq \frac{K + b}{K} \epsilon$

[LLB, Laska, Boufounos, Baraniuk, 13] ($b = 0$)
[LLD, Degraux, De Vleeschouwer, 13] ($b \neq 0$)

If $M = O(\epsilon^{-2} K \log N) \rightarrow$ Binary $\epsilon$ Stable Embedding (B$\epsilon$SE):

\[ d_{\text{ang}}(x, s) - \epsilon \leq d_H(A(x), A(s)) \leq d_{\text{ang}}(x, s) + \epsilon \]

$\neq$ RIP like

[LLB, Laska, Boufounos, Baraniuk, 13] [Plan, Vershynin, 11, 12]

Lesson: possible recovery should be as consistent as possible.
**Easiest 1-bit reconstruction**

**Fact:** If \( M = O(\epsilon^{-2} K \log N / K) \) (for \( x \in \Sigma_K \) fixed, \( \forall s \in \Sigma_K \))

\[
|\frac{\sqrt{\pi}/2}{M} \langle \text{sign}(\Phi x), \Phi s \rangle - \langle x, s \rangle| \leq \epsilon \quad \text{(w.h.p)}
\]

[Plan, Vershynin, 12]

\[
\downarrow
\]

Let \( x \in \Sigma_K \cap S^{N-1} \) and \( q = \text{sign}(\Phi x) \).

Compute

\[
\hat{x} = \frac{\pi}{2M} \mathcal{H}_K(\Phi^* q)
\]

Then, if previous property holds, \( \|x - \hat{x}\| \leq 2\epsilon \).

[LJ, Degraux, De Vleeschouwer, 13]

\[
\propto \arg \max_{u \in \mathbb{R}^N} q^T \Phi u \quad \text{s.t.} \quad \|u\|_0 \leq K, \quad \|u\| = 1
\]

≡ “PV-L0 problem”

[Bahmani, Boufounos, Raj, 13]
Binary Iterative Hard Thresholding (BIHT)

Idea: “greedily”

- forcing consistency
- forcing sparsity

\[ J(x') = 0 \quad \text{with} \quad (\lambda)_- = (\lambda - |\lambda|)/2 \]

or

\[ J(x') = \| [\text{diag}(q)(\Phi x')]_\|_1 \]
Binary Iterative Hard Thresholding (BIHT)

Idea: “greedily”

- forcing consistency
- forcing sparsity

\[ J(x') = \sum_{j=1}^{M} \left| \left( \text{sign}(\langle \varphi_j, x \rangle) \langle \varphi_j, x' \rangle) \right)_- \right| \]

with \( \lambda_- = (\lambda - |\lambda|)/2 \)

or \( J(x') = \| \left[ \text{diag}(q)(\Phi x') \right]_- \|_1 \)

Objective: arg\(\min_u J(u) \) s.t. \( \|u\|_0 \leq K \)

Given \( q = A(x) \) and \( K \), set \( l = 0, x^0 = 0 \):

\[
\begin{align*}
\alpha^{l+1} &= x^l + \frac{\tau}{2} \Phi^T (q - A(x^l)), \\
x^{l+1} &= \mathcal{H}_K(\alpha^{l+1}), \quad l \leftarrow l + 1
\end{align*}
\]

(“gradient” towards consistency)

(\( \tau > 0 \) controls gradient descent)

(proj. \( K \)-sparse signal set)
3. Bridging 1-bit & $B$-bit CS?
Bridging 1-bit & $B$-bit CS?

- $B$-bit quantizer defined with thresholds:

\[
\lambda \in \mathcal{R}_i = [\tau_i, \tau_{i+1}) \iff (\text{sign}(\lambda - \tau_i) = +1 \& \text{sign}(\lambda - \tau_{i+1}) = -1)
\]

- Can we combine multiple thresholds in 1-bit CS?

\[
q_i = Q[\lambda]
\]
Bridging 1-bit & $B$-bit CS?

Given $\mathcal{T} = \{\tau_j\}$ and $\Omega = \{q_j\}$ ($|\mathcal{T}| = 2^B + 1 = |\Omega| + 1$), let’s define

$$J(\nu, \lambda) = \sum_{j=2}^{2^B} w_j \left| (\text{sign} (\lambda - \tau_j) (\nu - \tau_j))_+ \right|,$$

with $w_j = q_j - q_{j-1}$.

Remark: for symmetric $Q$, $Q(\lambda) = \sum_{j=2}^{2^B} w_j \text{sign} (\lambda - \tau_j)$
Bridging 1-bit & $B$-bit CS?

Given $\mathcal{T} = \{\tau_j\}$ and $\Omega = \{q_j\}$ ($|\mathcal{T}| = 2^B + 1 = |\Omega| + 1$), let’s define

$$J(\nu, \lambda) = \sum_{j=2}^{2^B} w_j \left| \left( \text{sign}(\lambda - \tau_j) (\nu - \tau_j) \right)_+ \right|,$$

with $w_j = q_j - q_{j-1}$.

**Illustration:** $\lambda \in [\tau_{j-1}, \tau_j)$, $\nu \in [\tau_j, \tau_j+1)$

“delocalized” BIHT $\ell_1$-sided norm

$$J(\nu, \lambda) = \left| \left( \text{sign}(\lambda - \tau_j) (\nu - \tau_j) \right)_+ \right| = (\nu - \tau_j)$$

(for $w_j = 1$)
Bridging 1-bit & $B$-bit CS?

Given $T = \{\tau_j\}$ and $\Omega = \{q_j\}$ ($|T| = 2^B + 1 = |\Omega| + 1$), let’s define

$$J(\nu, \lambda) = \sum_{j=2}^{2^B} w_j \left| \left( \text{sign} (\lambda - \tau_j) (\nu - \tau_j) \right)_+ \right|,$$

with $w_j = q_j - q_{j-1}$.

Illustration: $\lambda \in [\tau_{j-1}, \tau_j)$, $\nu \in [\tau_{j+1}, \tau_{j+2})$

$$J(\nu, \lambda) = (\nu - \tau_j) + (\nu - \tau_{j+1})$$

(for $w_j = 1$)
Bridging 1-bit & $B$-bit CS?

Given $\mathcal{T} = \{\tau_j\}$ and $\Omega = \{q_j\}$ ($|\mathcal{T}| = 2^B + 1 = |\Omega| + 1$), let’s define

$$J(\nu, \lambda) = \sum_{j=2}^{2^B} w_j \left| (\text{sign}(\lambda - \tau_j)(\nu - \tau_j))_\pm \right|,$$

with $w_j = q_j - q_{j-1}$.

Illustration:

$$(\nu - \tau_j) + (\nu - \tau_{j+1})$$

(for $w_j = 1$)
Bridging 1-bit & $B$-bit CS?

Given $\mathcal{T} = \{\tau_j\}$ and $\Omega = \{q_j\}$ ($|\mathcal{T}| = 2^B + 1 = |\Omega| + 1$), let's define

$$J(\nu, \lambda) = \sum_{j=2}^{2^B} w_j \left| (\text{sign} (\lambda - \tau_j) (\nu - \tau_j))_+ \right|,\n$$

with $w_j = q_j - q_{j-1}$.

Illustration: more bins, $\lambda \in \mathcal{R}_5$
Bridging 1-bit & $B$-bit CS?

Given $T = \{\tau_j\}$ and $\Omega = \{q_j\}$ ($|T| = 2^B + 1 = |\Omega| + 1$), let’s define

$$J(\nu, \lambda) = \sum_{j=2}^{2^B} w_j \left| (\text{sign} (\lambda - \tau_j)(\nu - \tau_j))_+ \right|,$$

with $w_j = q_j - q_{j-1}$.

For $\mathbf{u}, \mathbf{v} \in \mathbb{R}^M$: $J(\mathbf{u}, \mathbf{v}) := \sum_{k=1}^{M} J(u_k, v_k)$ (component wise)

Remarks:

- $J$ is convex in $\nu$
- For $B = 1$ ($j = 2$ only):
  $$J(\mathbf{u}, \mathbf{v}) \propto \|(\text{sign} (\mathbf{v}) \odot \mathbf{u})_+\|_1 \rightarrow \ell_1$-sided 1-bit energy
- For $B \gg 1$:
  $$J(\nu, \lambda) \rightarrow \frac{1}{2}(\nu - \lambda)^2 \text{ and } J(\mathbf{u}, \mathbf{v}) \rightarrow \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|^2 \text{ (quadratic energy)}$$
Bridging 1-bit & $B$-bit CS?

- Let’s define an *inconsistency* energy:
  
  \[ \mathcal{E}_B(u) := J(\Phi u, q) \]
  
  with \( q = Q_B[\Phi x] \) and \( \mathcal{E}_B(x) = 0 \)

- Idea: Minimize it in \( \Sigma_K \) (as for Iterative Hard Thresholding)

[Blumensath, Davies, 08]

\[ \min_{u \in \mathbb{R}^N} \mathcal{E}_B(u) \text{ s.t. } \|u\|_0 \leq K, \]
Bridging 1-bit & $B$-bit CS?

- Let’s define an *inconsistency* energy:

$$
\mathcal{E}_B(u) := J(\Phi u, q) \text{ with } q = Q_B[\Phi x] \text{ and } \mathcal{E}_B(x) = 0
$$

- Idea: Minimize it in $\Sigma_K$ (as for Iterative Hard Thresholding)

$$
\min_{u \in \mathbb{R}^N} \mathcal{E}_B(u) \text{ s.t. } \|u\|_0 \leq K,
$$

- combinatorial but greedy solution (as for IHT):

$$
x^{(n+1)} = H_K[x^{(n)} - \mu \partial \mathcal{E}_B(x^{(n)})] \text{ and } x^{(0)} = 0. \quad (\mu > 0, \text{ see after})$$

\[\Phi^*(\text{sign } (\Phi u) - \text{sign } (\Phi x))\] $$\Phi^*(Q_B(\Phi u) - q)$$  \[\text{BIHT!}\]  \[\text{Quantized IHT (QIHT)}\]  \[\text{IHT!}\]  “all that.. for this!” ;-)

[Blumensath, Davies, 08]
Bridging 1-bit & $B$-bit CS?

- So, QIHT reads
  \[ x^{(n+1)} = \mathcal{H}_K \left[ x^{(n)} + \mu \Phi^* (q - Q_B(\Phi x^{(n)})) \right] \] and $x^{(0)} = 0$.

- Setting $\mu$?
  - for $B = 1$, $\mu$ has no impact
  - for high $B$, IHT solution: [Blumensath, Davies, 08]
    \[
    \text{If } \Phi / \sqrt{M} \text{ is RIP, } \mu < \frac{1}{(1 + \delta_{2K}) M}
    \]
  - for any $B$, heuristic: $\mu = \frac{1}{M} \left( 1 - \sqrt{cK/M} \right)$ for some $c > 0$.

Validated by extensive simulations ($c = 3$)

SNR: [Graphs showing performance with varying $\mu$ and $b$]
QIHT simulations

\(N = 1024, \ K = 16, \ R = BM \in \{64, 128, \cdots, 1280\}, 100\) trials (+ Lloyd-Max Gauss. Q.)

Note: entropy could be computed instead of \(B\) (e.g., for further efficient coding)

\*: almost “6dB per bit” gain
QIHT simulations

\(N = 1024, \ K = 16, \ R = BM \in \{64, 128, \cdots, 1280\}, 100 \text{ trials} \ (\pm \text{Lloyd-Max Gauss. Q.})\)

Note: entropy could be computed instead of \(B\) (e.g., for further efficient coding)

QIHT stops if

\[
\frac{\|x^{(n+1)} - x^{(n)}\|}{\|x^{(n)}\|} < 10^{-4}
\]

or \(n = 500\).

*: almost “6dB per bit” gain
**QIHT simulations**

\[ N = 1024, \ K = 16, \ R = BM \in \{64, 128, \cdots, 1280\}, \ 100 \text{ trials} (+ \text{Lloyd-Max Gauss. Q.}) \]

Note: entropy could be computed instead of \( B \) (e.g., for further efficient coding)

**Running time:**

![Graphs showing running time for QIHT, IHT, and BPDN](image)

**Consistency?**

![Graphs showing consistency for QIHT, IHT, and BPDN](image)
Conclusions

- IHT framework can integrate quantization
- Lot of theoretical works to do/come...
  - BIHT/QIHT convergence/optimality guarantees?
  - New embedding results?
    
    \[
    Q(\lambda) = \sum_{j=2}^{2^B} w_j \text{sign}(\lambda - \tau_j) = \text{sum of 1-bit costs + thresholds}
    \]
  - Blend of RIP/BεSE? as in
    \[
    (1 - \delta)\|x - x'\| - \epsilon \leq \text{dist}(Q(\Phi x), Q(\Phi x')) \leq (1 + \delta)\|x - x'\| + \epsilon
    \]
    for sparse \(x, x'\) and with \(\epsilon \to B, M 0\) and \(\delta \to M 0\)?
- Coming:
  - Application to \(B\)-bit compressive sensors (RMPI)
Further Reading


