Convex Rate-Distortion Formulation for Mixed-Fidelity Compression

Marc Nasser
Department of Electrical Engineering, Stanford University
mnasser@stanford.edu

December 6, 2025

1 Introduction

Lossy compression systems balance the competing objectives of minimizing bitrate and maintaining acceptable distortion. However, standard approaches, including Lloyd-Max quantization and modern neural compressors, are inherently nonconvex, provide no global optimality guarantees, and offer limited mechanisms for imposing heterogeneous fidelity requirements across different regions of the source distribution (1; 2).

Many practical settings demand *mixed-fidelity* compression: certain subsets of the data, such as safety-critical regions or metadata, require significantly lower distortion than the rest. Enforcing such locally tighter distortion bounds is sometimes difficult in traditional nonconvex frameworks, as they do not provide explicit control over it. It also becomes unclear whether the resulting encoder is anywhere near the fundamental rate-distortion limit.

In this project, we formulate entropy-constrained compression as a convex optimization problem over stochastic encoders. For a fixed reconstruction codebook, we minimize the mutual information $I(X;Y)$ (a convex objective) subject to a global distortion constraint and optional subset-specific distortion constraints. This convex perspective provides globally optimal stochastic encoders for a fixed codebook, exposes dual variables that quantify the marginal 'price' of distortion, and explicitly accommodates mixed-fidelity requirements.
By alternating convex encoder updates with simple codebook refinements and support merging, we obtain practical compressors whose empirical rate closely approaches the convex lower bound.

This report develops the formulation, describes the optimization pipeline, and evaluates its performance on several one and two dimensional sources, with mixed-fidelity constraints. We also compare our solutions with Lloyd-max, illustrating the fundamental differences between the two algorithms.

2 Literature Review

The foundation of this work lies in the classical theory of rate-distortion for discrete memoryless sources. The rate-distortion function $R(D)$ characterizes the minimum achievable rate for representing a source X under an average distortion constraint, and is typically computed through iterative algorithms that optimize over a conditional distribution $Q(y|x)$ mapping source symbols to reconstruction symbols. The most influential class of methods is the Blahut-Arimoto family of algorithms (1; 2), which alternately update the encoder Q and the output marginal p_Y to minimize the mutual information $I(X;Y)$ subject to a distortion constraint.

Although the Blahut-Arimoto procedure implicitly leverages the convexity of mutual information in the channel law, it is traditionally presented in a single distortion constraint setting and does not provide a direct mechanism for handling heterogeneous fidelity specifications. Moreover, while these algorithms converge to stationary points of the rate-distortion curve, they do not expose dual information that quantifies the marginal 'cost' of tightening or relaxing distortion constraints. Such dual variables are valuable when one wishes to interpret the trade-offs inherent in mixed-fidelity compression.

Recent advances in lossy compression using neural networks or improved variants of Lloyd–Max quantization demonstrate strong empirical performance, but they inherit the fundamental nonconvexity

of the underlying optimization problems. As a result, they lack global optimality guarantees and offer little interpretability regarding how far a practical encoder is from the theoretical rate-distortion limit. Additionally, these approaches are not designed to impose region-specific fidelity requirements, making them ill-suited for mixed-fidelity compression scenarios found in applications such as safety-critical sensing or heterogeneous-importance data streams.

In contrast, convex formulations allow one to obtain globally optimal stochastic encoders for a fixed reconstruction alphabet and to incorporate additional affine constraints without altering convexity. We get several benefits: (i) the ability to certify lower bounds on achievable rates; (ii) access to dual variables that quantify the marginal price of distortion; and (iii) a principled foundation for heuristic extensions such as codebook refinement and merging. These attributes make convex optimization a compelling tool for understanding and designing mixed-fidelity lossy compressors.

3 Methods

Our approach alternates three components. (i) Solving a convex rate-distortion program over the stochastic encoder Q, for a fixed codebook. (ii) Refining the reconstruction codebook Y via gradient-based updates. (iii) Merging low-mass or redundant codewords to reduce support size. Each step is formulated to preserve convexity and to support mixed-fidelity constraints.

3.1 Optimization Pipeline

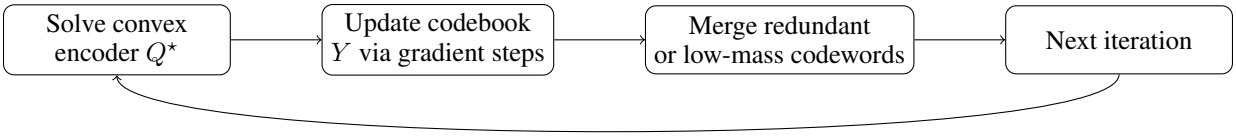

Figure 1: High-level optimization cycle. Each block is detailed below.

3.2 Convex Core Program

For a discrete source X with empirical distribution $p_X(x)$ and a fixed reconstruction alphabet $Y = \{y_j\}_{j=1}^m$, we parameterize the encoder by a conditional distribution

$$Q = (Q_{j|x})_{x,j}, \qquad Q_{j|x} = P(Y = y_j \mid X = x).$$

The rate surrogate is the mutual information

$$I(X;Y) = \sum_x p_X(x) \sum_j Q_{j|x} \log\left(\frac{Q_{j|x}}{p_Y(j)}\right),$$

which is jointly convex in the stochastic matrix Q when written in this relative-entropy form. The marginal p_Y is given by $p_Y(j) = \sum_x p_X(x)\, Q_{j|x}$.

We impose two distortion constraints: (i) a global average distortion limit D, and (ii) an optional mixed-fidelity constraint with a tighter distortion tolerance D' on a subset $S \subseteq \mathcal{X}$.

Let $d(x, y_j)$ denote the distortion measure (here, squared error). The optimization problem is:

$$\min_{Q \geq 0} I(X;Y) \quad \text{s.t.} \sum_j Q_{j|x} = 1 \,\forall x,\ \sum_{x,j} p_X(x) Q_{j|x} d(x,y_j) \leq D,\ \sum_{x \in S, j} p_X(x) Q_{j|x} d(x,y_j) \leq D',\ p_Y(j) = \sum_x p_X(x) Q_{j|x}$$

This is a convex relative-entropy program: - the objective is convex in Q, - all constraints are affine, - the solution is the globally optimal stochastic encoder for a given fixed codebook.

The dual variables corresponding to the distortion constraints quantify the marginal price of reducing distortion globally or within the fidelity subset, offering interpretable sensitivity information.

3.3 Codebook Updates via Gradient Descent

After solving for $Q^\star$, we refine the reconstruction points by minimizing the expected distortion conditioned on each output symbol. Each codeword y_j is updated via

$$y_j \leftarrow y_j - \eta \nabla_{y_j} \mathbb{E}[\, d(X, y_j) \mid Y = j \,],$$

where $\eta > 0$ is a step size. Using the law of total expectation,

$$\mathbb{E}[d(X, y_j) \mid Y = j] = \frac{1}{p_Y(j)} \sum_x p_X(x) Q_{j|x} d(x, y_j).$$

For squared-error distortion $d(x, y_j) = \|x - y_j\|_2^2$, the gradient becomes

$$\nabla_{y_j} d(x, y_j) = 2(y_j - x),$$

and the stationary point satisfies

$$y_j = \mathbb{E}[X \mid Y = j] = \frac{\sum_x p_X(x) Q_{j|x} x}{\sum_x p_X(x) Q_{j|x}}.$$

Thus, for MSE the update reduces to a centroid computation, weighted by the stochastic encoder $Q^\star$ rather than hard assignments.

3.4 Codeword Merging and Support Reduction

To remove redundant reconstruction points and encourage compact codebooks, we merge codewords when either:

- their marginal mass is negligible: $p_Y(j) < \varepsilon$, or
- they are nearly redundant in distortion: $d(y_j, y_k) < \delta$.

If either condition holds, codeword j is merged into k, and the encoder entries are updated by reassigning $Q_{j|x}$ to $Q_{k|x}$. This preserves total mass and keeps the optimization stable.

Such merging steps reduce the effective cardinality of Y, which in turn reduces the achievable entropy $H(Y)$ of the induced encoder. In practice this step is crucial for obtaining deployable low-rate compressors whose empirical rate $H(Y)$ stays close to the convex lower bound $I(X;Y)$ produced by the optimization.

The full optimization pipeline produces a sequence of increasingly compact and information-efficient compressors. In the next section, we evaluate this framework on several synthetic one and two-dimensional sources, examine its behavior under mixed-fidelity specifications, and compare its empirical rate-distortion performance against classical nonconvex baselines such as Lloyd-Max quantization.

4 Results and Analysis

We evaluate the convex rate-distortion framework on synthetic one and two-dimensional sources with squared-error distortion under both global and mixed-fidelity constraints. Our experiments assess: (i) the validity of the convex formulation via comparison with known Gaussian $R(D)$ curves (ii) the effect of mixed-fidelity constraints on the optimizer and resulting trade-offs; (iii) the behavior of the full encoder-codebook refinement pipeline and (iv) comparisons with classical nonconvex methods such as Lloyd-Max quantization. Overall, the convex approach produces globally optimal stochastic encoders for fixed codebooks, offers clear interpretability through dual sensitivities, and achieves empirical rates close to the convex lower bound while satisfying fidelity requirements. For each experiment, additional plots conveying the same message of the main results plots but on different sources X are shown in the Appendix 5

4.1 Validation of the Convex Core

We begin by validating the convex rate-distortion formulation on sources whose theoretical $R(D)$ behavior is known. Experiments are conducted on discretized one and two-dimensional Gaussian sources under squared-error distortion. As they convey the similar conclusion, we focus on just one source in this report.

1D Gaussian Source. Figure 2 (left) shows the empirical distribution of the discretized 1D Gaussian. Solving the convex program over varying distortion budgets produces the rate–distortion curve in Figure 2 (right), plotted against the closed-form Gaussian rate-distortion function

$$R_{\text{Gauss}}(D) = \tfrac{1}{2}\log\left(\frac{\sigma^2}{D}\right).$$

The close match between the discrete convex solution and the theoretical curve confirms both the correctness of the convex formulation and the stability of its numerical implementation. Very small deviations arise from discretization of the codebook (since the theoretical solution assumes a continuous one). Overall, these experiments verify that the convex core reliably reproduces optimal rate-distortion behavior.

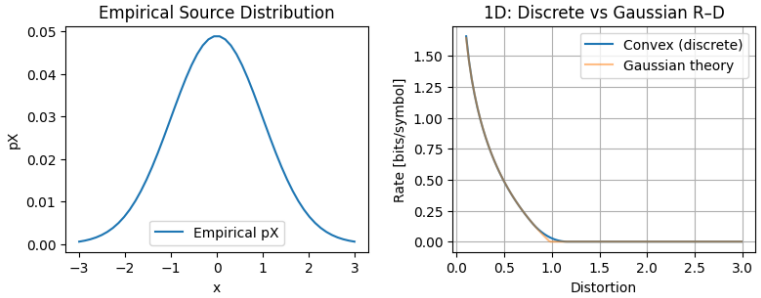

Figure 2: Left: empirical source distribution for the discretized 1D Gaussian. Right: resulting convex rate-distortion curve compared with the theoretical Gaussian $R(D)$.

4.2 Mixed-Fidelity Constraints and Dual Variables

Next, we show an example of how having access to dual variables help us better understand our solution and the effects of the constraints on our codebook. Here, we demonstrate the effect of imposing a tighter distortion budget on a high-probability subset S of a 1D Gaussian source. By respecting the tighter bound for the mixed-fidelity subset, the encoder works on preserving accuracy in the targeted region even when the global budget would allow more aggressive compression.

To understand how this constraint shapes the solution, we track the dual variables associated with the global and subset-specific distortion constraints, together with the resulting rate–distortion curve. Figure 3 shows both quantities. The subset dual variable λ_{subset} is an order of magnitude larger than λ_{global} across the range of distortions, indicating that the high-fidelity subset dominates the rate cost. Additionally, as the global distortion constraint is relaxed, λ_{global} decreases, while λ_{subset} increases, since the burden of limiting rate reduction shifts toward the tighter subset constraint on S, which therefore carries a higher marginal 'price'. Correspondingly, the mixed-fidelity $R(D)$ curve is noticeably flatter than the unconstrained Gaussian curve observed in Figure 2: the encoder cannot collapse its support as aggressively without violating the stricter constraint on S, leading to a higher achievable rate for a given global distortion.

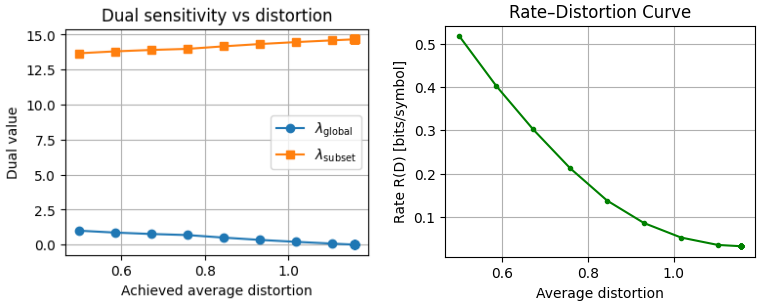

Figure 3: Left: dual variables for global and subset distortion constraints. Right: mixed-fidelity rate–distortion curve. The tighter subset constraint makes the $R(D)$ curve shallower and prevents aggressive rate reduction.

Overall, these results confirm that mixed-fidelity constraints are handled cleanly by the convex formulation and significantly influence the achievable rate–distortion trade-off. It also demonstrates how having a convex program is useful in understanding the final codebook, and how constraints affect the minimizer and optimal encoder.

4.3 Iterative Refinement

Then, we evaluate our whole algorithm pipeline described in Methods.

Figure 4 shows the initial and refined codebooks for a multi-gaussian 1D source with three separated modes. The initial codebook places many redundant points across, while the refined codebook concentrates its support near the source modes, as expected for squared-error distortion.

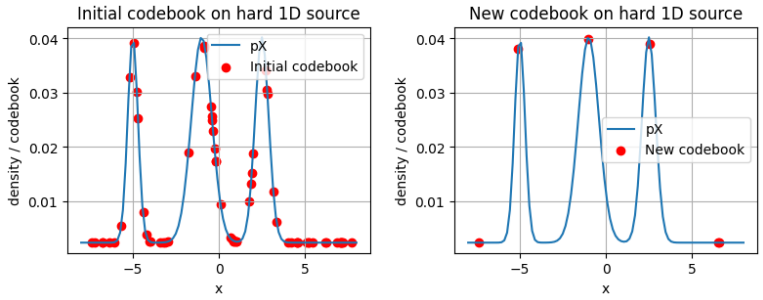

Figure 4: Left: initial codebook for a hard 1D trimodal source. Right: refined codebook after a few iterations of alternating convex encoder updates and codebook refinement.

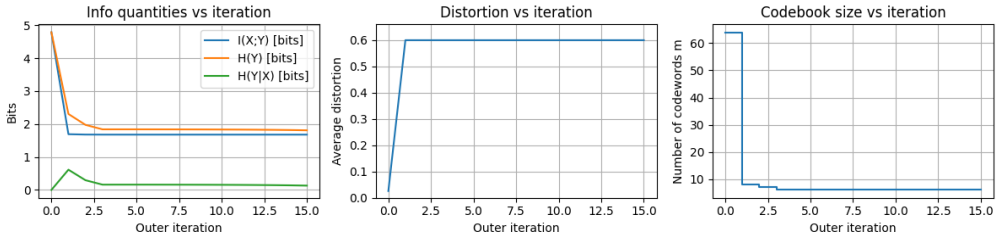

Figure 5: Evolution of $I(X;Y)$, $H(Y)$, $H(Y|X)$, average distortion, and codebook size over outer iterations.

We also demonstrate refinement on a 2D mixed-fidelity source with four high-density regions and a central fidelity constraint. As shown in Figure 6, the refined codebook aligns precisely with the source modes while respecting the tighter distortion requirement on the central region, as well as being centered by E[X], illustrating that the refinement process adapts naturally to multidimensional and mixed-fidelity settings.

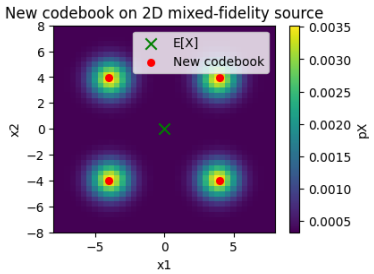

Figure 6: Refined codebook on a 2D mixed-fidelity source. Reconstruction points align with the source modes while respecting the high-fidelity region at the center.

The evolution of information quantities, distortion, and codebook size across outer iterations is shown in Figure 5. Mutual information $I(X;Y)$ and entropy $H(Y)$ drop sharply during the initial iterations as redundant codewords are merged. The codebook size falls from over sixty points to five, and average distortion converges to its upper bound to allow minimization of rate. We quickly converge to our solution.

4.4 Comparison with Lloyd–Max Quantization

We conclude by comparing our convex approach to the classical Lloyd-Max quantizer on the 1D trimodal source. For a matched codebook size of $m = 5$, Figure 7 shows the resulting reconstruction points. Both methods place codewords near the three source modes, but are not exactly the same. This difference in final behavior reflects the differing optimization objectives: Lloyd-Max minimizes distortion, while our method minimizes mutual information.

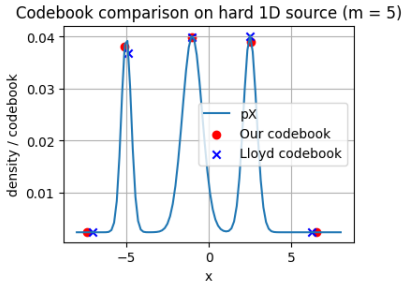

Figure 7: Comparison of reconstruction points for our method and Lloyd–Max on a hard 1D source with $m = 5$.

The quantitative results reflect this distinction. Our method achieves a substantially lower mutual information and output entropy:

$$I(X;Y) = 1.6777 \text{ bits}, \qquad H(Y) = 1.7969 \text{ bits}, \qquad D = 0.6000,$$

while Lloyd–Max achieves lower distortion but at a higher rate:

$$I(X;Y) = 1.9099 \text{ bits}, \qquad H(Y) = 1.9099 \text{ bits}, \qquad D = 0.4170.$$

These results illustrate the trade-off between the two objectives: our convex formulation pushes for minimum rate by pushing distortion to its upper limit, while Lloyd–Max favors distortion at the cost of a higher empirical rate. This distinction also highlights the additional control our algorithm provides over Lloyd-Max: we control distortion limits while minimizing rate, but Lloyd-Max doesn't control rate limits while minimizing distortion.

5 Conclusion

We introduced a convex rate-distortion framework that minimizes mutual information under global and mixed-fidelity distortion constraints. The convex formulation allows explicit mixed-fidelity constraints formulations, yields globally optimal stochastic encoders for fixed codebooks, and provides dual sensitivities that make fidelity trade-offs transparent and interpretable, allowing more control on the settings of the codebook. Combined with simple codebook refinement and merging, the method produces compact encoders whose empirical rates track the convex lower bound.

Experiments on 1D and 2D sources confirmed the correctness of the convex core and illustrated how mixed-fidelity constraints are interpretable and shape the achievable rate. Compared with Lloyd-Max, our approach achieves lower rates at the cost of higher distortion, reflecting its rate-minimizing objective. More importantly, the convex formulation offers explicit control over distortion limits and rate efficiency that most quantizers cannot express.

References

References

[1] R. E. Blahut, "Computation of channel capacity and rate–distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[2] S. Arimoto, "An algorithm for calculating the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.

A Appendix / Supplemental Material

Additional experiments conveying the same message as the main results section.

A.1 2D Solver Convexity Verification

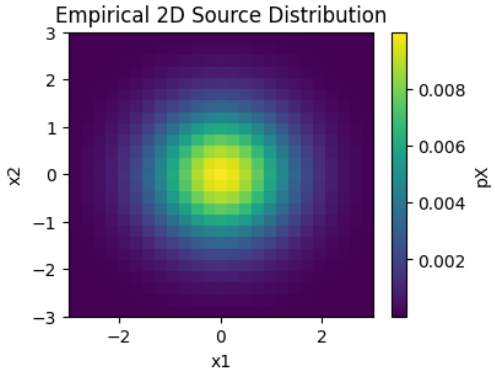

Figure 8: 2D Gaussian distribution.

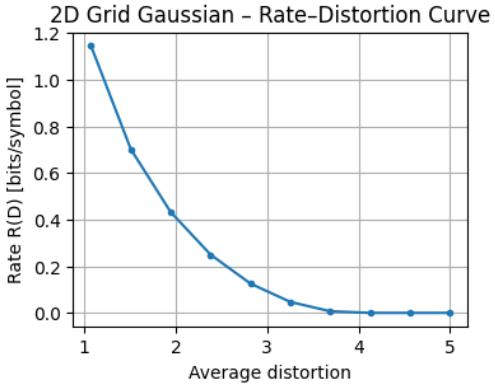

Figure 9: Rate–distortion curve verifying convexity of the 2D solver.

A.2 Iterative Refinement on Mixed-Distribution Source

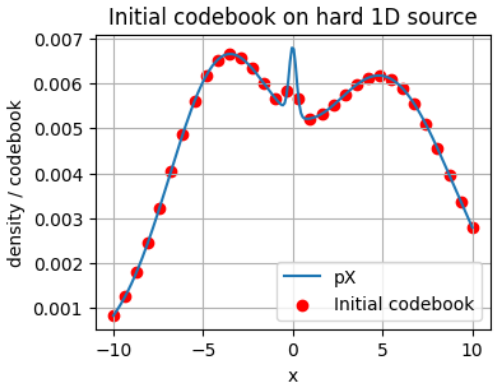

Figure 10: Initial codebook on a mixed-distribution source.

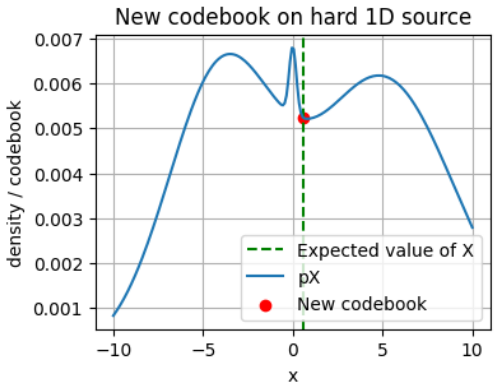

Figure 11: Refined codebook after iterative optimization.

A.3 Comparisons with Lloyd–Max

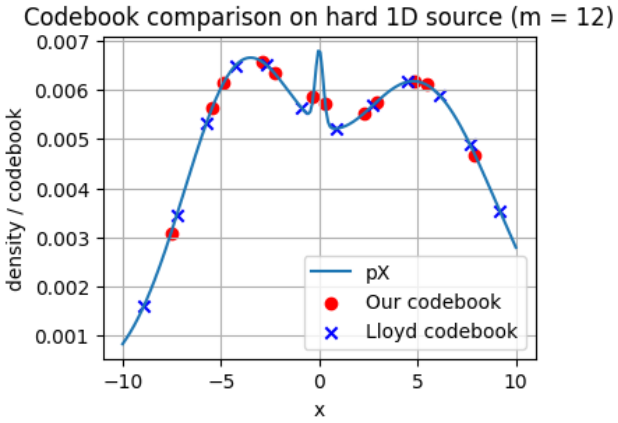

Figure 12: Comparison of final codebooks: Lloyd–Max vs. our method.