Continuous learning with errors

Replace residues by fixed-precision reals: division by q moves retained residue bits to right of binary point. For $q = 2^k$, $x \bmod q = \sum_{i=0}^{k-1} b_i 2^i$ and $(x \bmod q)/q = \sum_{i=0}^{k-1} b_i 2^{i-k}$.

Definition 1 (Continuous LWE (CLWE))**.** *Fix $\gamma, \beta > 0$ and a hidden unit vector $\mathbf{w} \leftarrow S^{n-1}$. Independent samples:*

$$\mathbf{a} \leftarrow N(0, I_n), \qquad b = \gamma \mathbf{a}^\top \mathbf{w} + e \pmod 1, \qquad e \leftarrow N(0, \beta^2).$$

- Search: recover $\mathbf{w}$; decision: distinguish from $(\mathbf{a}, u)$ with independent $u \leftarrow [0, 1)$ uniform.
- Without mod 1: ordinary regression.
- Along $\mathbf{w}$: label bands repeat with period $1/\gamma$, blurred by noise e of width β.

Theorem 1 (Hardness of CLWE)**.** *A polynomial-time algorithm for* decisional CLWE *gives a polynomial-time algorithm for* decisional LWE*.*

- Conjectured hard for $\gamma \gtrsim \sqrt{n}$, $\beta \in (0, 1)$, and $\gamma/\beta \leq \text{poly}(n)$.
- Near-zero noise ($\beta \approx 0$) permits polynomial-time lattice-reduction attacks.

Homogeneous CLWE

Recall HW1. For $\mathbf{a} \sim N(0, I_n)$ and unit $\mathbf{w}$,

$$\mathbf{a} = t\mathbf{w} + \mathbf{z}, \quad t = \mathbf{a}^\top \mathbf{w}, \quad \mathbf{z} \sim N(0, I_n - \mathbf{w}\mathbf{w}^\top), \quad t \perp \mathbf{z}.$$

- HW1: keep $|t| \leq b$ (one slab); for the sketch set $b = \beta/\gamma$.
- CLWE: $b = (\gamma t + e) \bmod 1$. Keep $b \approx 0$, erase b: $\gamma t + e \approx k$, $k \in \mathbb{Z}$ (spacing $1/\gamma$, width $\sim \beta/\gamma$).
- In both cases selection depends on t (and independent e), not $\mathbf{z}$. Hence $\mathbf{z}$ is *unchanged*.



$|t| \leq b$ spacing $1/\gamma$, width $\sim \beta/\gamma$

Sketch: z_1 is one coordinate in $\mathbf{w}^\perp$; dashed circles show original Gaussian contours; gray bands indicate horizontal width (not a hard cutoff for hCLWE).

Definition 2 (Homogeneous CLWE (hCLWE))**.** *Given m unlabeled vectors $\mathbf{a}_i \in \mathbb{R}^n$, distinguish:*

- Null: *independent* $\mathbf{a}_i \leftarrow N(0, I_n)$.
- Planted: *one hidden unit* $\mathbf{w}$*; Gaussian in every direction orthogonal to* $\mathbf{w}$*, while* $a_i^{\parallel} = \mathbf{a}_i^\top \mathbf{w}$ *lies near* k/γ*,* $k \in \mathbb{Z}$*, with peak width about* β/γ*.*

- Along known $\mathbf{w}$: periodic spikes are easy to detect.
- With unknown $\mathbf{w}$: the task is to find their direction.

Proposition 1 (CLWE $\to$ hCLWE)**.** *An efficient hCLWE distinguisher yields an efficient decisional CLWE distinguisher (with suitable parameters).*

Proof sketch. Given $(\mathbf{a}, b)$, retain samples with $b \approx 0 \pmod 1$; then erase b.

- *Null:* b is independent uniform, so retained $\mathbf{a} \sim N(0, I_n)$.
- *CLWE:* $\gamma t + e \approx k$ for some $k \in \mathbb{Z}$: periodic Gaussian slabs along $\mathbf{w}$. By HW1, $\mathbf{z}$ stays independent and Gaussian.

A suitable Gaussian filter gives hCLWE with adjusted noise and retains an inverse-polynomial fraction of samples. □

A Gaussian mixture draws from one of several Gaussian components, chosen according to fixed mixing weights. Idealizing $b \approx 0$ as conditioning on $b = 0$, the planted law is

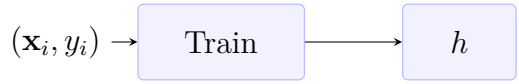
$$P_{\mathbf{w}} = \sum_{k \in \mathbb{Z}} \pi_k \, N(\mu_k \mathbf{w}, \Sigma), \quad \mu_k = \frac{\gamma k}{\gamma^2 + \beta^2},$$
$$\Sigma = I_n - \frac{\gamma^2}{\gamma^2 + \beta^2} \mathbf{w}\mathbf{w}^\top, \quad \pi_k \propto \exp\left(-\frac{k^2}{2(\gamma^2 + \beta^2)} \right).$$

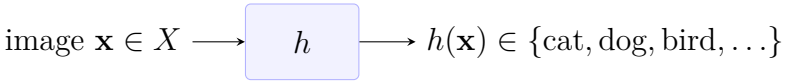
Proposition 2. *Assuming hCLWE is hard, there are Gaussian mixtures that are hard to distinguish from a standard Gaussian in polynomial time.*

Write $\mathbf{a} = a^{\parallel}\mathbf{w} + \mathbf{a}^{\perp}$. For fixed integer k, the band $\gamma a^{\parallel} + e \approx k$ makes $a^{\parallel}$ approximately Gaussian, centered near k/γ with width β/γ; $\mathbf{a}^{\perp}$ remains an independent standard Gaussian.

Thus each k gives an approximately Gaussian component centered near $(k/\gamma)\mathbf{w}$, narrow along $\mathbf{w}$ but standard in orthogonal directions. Mixing with the probabilities of each k gives the planted distribution; the null is one standard Gaussian.

Adversarial Robustness in Classification





Accurate predictions on ordinary inputs need not survive deliberately chosen changes.

- *Sensitivity:* small, label-preserving change → different prediction.
- *Invariance:* different correct label → same prediction.

Examples:

- Stop-sign stickers or reworded spam fool a classifier (sensitivity).
- Attacker turns a benign email into phishing; filter still predicts "benign" (invariance).

Sensitivity attacks

Definition 3 (Classification problem)**.** *A classification problem consists of an input set X, a finite label set $Y = \{1, \ldots, K\}$, and a distribution D over labeled examples $(x, y) \in X \times Y$. A classifier is a function $h : X \to Y$. Its classification error is*

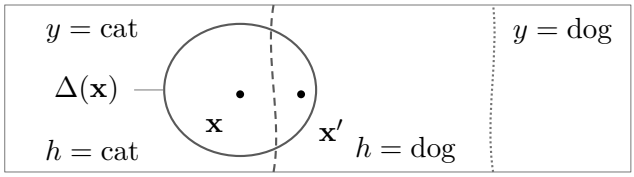
$$\Pr_{(x,y)\sim D}[h(x) \neq y].$$

- Examples: images; spam; diagnosis; content matching.
- Sensitivity = adversarial example / evasion (fake-nose example).

Definition 4 (Sensitivity attack)**.** *Write $y(x)$ for the correct label of x, and let $\Delta(x) \subseteq X$ be the allowed perturbations of x. A sensitivity attack A succeeds on a correctly classified input x if $A(x) \in \Delta(x)$, $y(A(x)) = y(x)$, and $h(A(x)) \neq h(x)$. Its success rate is*

$$\Pr_{(x,y)\sim D}\Big[A(x) \text{ succeeds} \mid h(x) = y\Big].$$

- Conditional success: *new* errors, not pre-existing errors.
- Common threat model: $\Delta(\mathbf{x}) = \{\mathbf{x}' : d(\mathbf{x}, \mathbf{x}') \leq \varepsilon\}$.
- $\mathbf{x}' \in \Delta(\mathbf{x}) \not\Rightarrow y(\mathbf{x}') = y(\mathbf{x})$.



Sensitivity attack (2D schematic): dashed = boundary of h; dotted = true-label boundary. Here $\mathbf{x}' \in \Delta(\mathbf{x})$ and $y(\mathbf{x}') = y(\mathbf{x})$, but $h(\mathbf{x}') \neq h(\mathbf{x})$.

- ℓ_∞: largest coordinate change; ℓ_2: aggregate change; ℓ_0: number of changed coordinates.

$$\|\mathbf{v}\|_\infty \le \|\mathbf{v}\|_2 \le \sqrt{n}\,\|\mathbf{v}\|_\infty.$$

$$\{\mathbf{x}' : \|\mathbf{x}' - \mathbf{x}\|_\infty \le r/\sqrt{n}\} \subseteq \{\mathbf{x}' : \|\mathbf{x}' - \mathbf{x}\|_2 \le r\}.$$

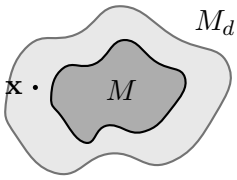
- Images: semantic manifold vs. pixel norms; metric only a proxy.
- Vision embeddings: possible implicit semantic geometry.

Accuracy does not imply robustness

- Uniform Hamming cube: high accuracy means $1 - \dfrac{|\{\mathbf{x} : h(\mathbf{x}) \neq y(\mathbf{x})\}|}{2^n} \approx 1$.
- Question: for most correctly classified $\mathbf{x}$, is every misclassified $\mathbf{x}'$ far from $\mathbf{x}$?

Theorem 2 (Nearby errors)**.** *For every fixed $\delta > 0$, any classifier on the Hamming cube with fixed positive error has an error within $O(\sqrt{n})$ bit flips of all but a δ fraction of inputs.*

- Error region: $M = \{\mathbf{x} : h(\mathbf{x}) \neq y(\mathbf{x})\}$; $|M|/2^n = \alpha$.
- $d_H(\mathbf{x}, M) = \min_{\mathbf{z} \in M} d_H(\mathbf{x}, \mathbf{z})$.
- $M_d = \{\mathbf{x} : d_H(\mathbf{x}, M) \le d\}$; $|M_d|/2^n =$ fraction near an error.



irregular error region

Claim 1 (Concentration of error regions)**.** *If $|M|/2^n = \alpha > 0$, then for every fixed $0 < \delta < 1$, there is $C_{\alpha,\delta} > 0$ such that $d \ge C_{\alpha,\delta}\sqrt{n}$ implies $|M_d| > (1 - \delta)2^n$.*

Proof. If M were a Hamming ball, its d-neighborhood would be obtained by increasing its radius by d. Concentration around $n/2$ would then give the conclusion for $d = O(\sqrt{n})$.

For arbitrary M, we need to know which shape minimizes the size of the d-neighborhood. This is given by Harper's inequality.

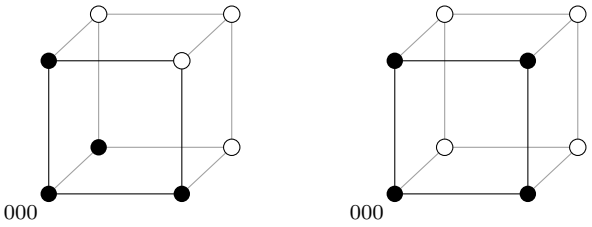
Harper's inequality. *Among sets of a given size, the d-neighborhood is smallest for a Hamming ball, possibly with only part of its outermost layer.*

Example: two sets of size 4 in $\{0,1\}^3$.

$$H = B_1(000) = \{000, 100, 010, 001\},$$
$$C = C_{\{1,2\}}(000) = \{000, 100, 010, 110\}.$$

$$|H| = |C| = 4.$$

Ball H **Subcube C**



Black dots are in the set; white dots are outside.

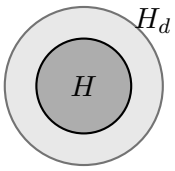
After one additional bit flip:

$$N_1(H) = \{0,1\}^3 \setminus \{111\}, \qquad |N_1(H)| = 7,$$
$$N_1(C) = \{0,1\}^3, \qquad |N_1(C)| = 8.$$

$$|N_1(H)| = 7 < 8 = |N_1(C)|.$$
$$|N_1(H) \setminus H| = 3 \quad \text{vs.} \quad |N_1(C) \setminus C| = 4.$$

In H, each new neighbor touches two marked vertices; in C, each touches one.

Let H be such a set, with $|H| = |M|$. Then



ball of equal size

$$|M| = |H| \qquad \Longrightarrow \qquad |M_d| \geq |H_d|.$$

2D schematic; Harper's inequality concerns Hamming neighborhoods on $\{0,1\}^n$.

Let $\mathbf{c}$ be the center of H and r its outer radius. The set H contains every string at distance less than r from $\mathbf{c}$. Consequently, H_d contains every string at distance less than $r + d$ from $\mathbf{c}$.

For uniform $\mathbf{x} \in \{0,1\}^n$,

$$d_H(\mathbf{x}, \mathbf{c}) \sim \text{Bin}(n, 1/2).$$

By Hoeffding's inequality,

$$\Pr\left[\left|d_H(\mathbf{x}, \mathbf{c}) - \frac{n}{2}\right| \geq t\right] \leq 2e^{-2t^2/n}.$$

Since $|H|/2^n = \alpha$, Hoeffding implies that its radius satisfies

$$r \geq \frac{n}{2} - O_\alpha(\sqrt{n}).$$

Similarly, all but a δ fraction of the cube lies within distance $\frac{n}{2} + O_\delta(\sqrt{n})$ of $\mathbf{c}$. Choose $d \geq C_{\alpha,\delta}\sqrt{n}$, with $C_{\alpha,\delta}$ sufficiently large. Then

$$r + d \geq \frac{n}{2} - O_\alpha(\sqrt{n}) + C_{\alpha,\delta}\sqrt{n} \geq \frac{n}{2} + O_\delta(\sqrt{n}).$$

Thus H_d contains more than a $1 - \delta$ fraction of the cube. By Harper's inequality,

$$|M_d| \geq |H_d| > (1 - \delta)2^n.$$

□

- Geometric *existence*, not an efficient attack.
- Nearby error $\not\Rightarrow$ permitted, label-preserving sensitivity attack.
- Example: three bit flips,

$$\mathbf{x} = (0, 1, 0, 1, 1, 0), \qquad \mathbf{x}' = (1, 1, 1, 1, 0, 0),$$
$$\mathbf{v} = \mathbf{x}' - \mathbf{x} = (1, 0, 1, 0, -1, 0).$$

$$\|\mathbf{v}\|_0 = \|\mathbf{v}\|_1 = 3, \qquad \|\mathbf{v}\|_2 = \sqrt{3}, \qquad \|\mathbf{v}\|_\infty = 1.$$

- In general, d bit flips give $\|\mathbf{v}\|_0 = \|\mathbf{v}\|_1 = d$, $\|\mathbf{v}\|_2 = \sqrt{d}$, $\|\mathbf{v}\|_\infty = 1$ (for $d \geq 1$).
- Real inputs: perturbation *magnitude* matters (high-/low-order bits).
- Related concentration for other product distributions.