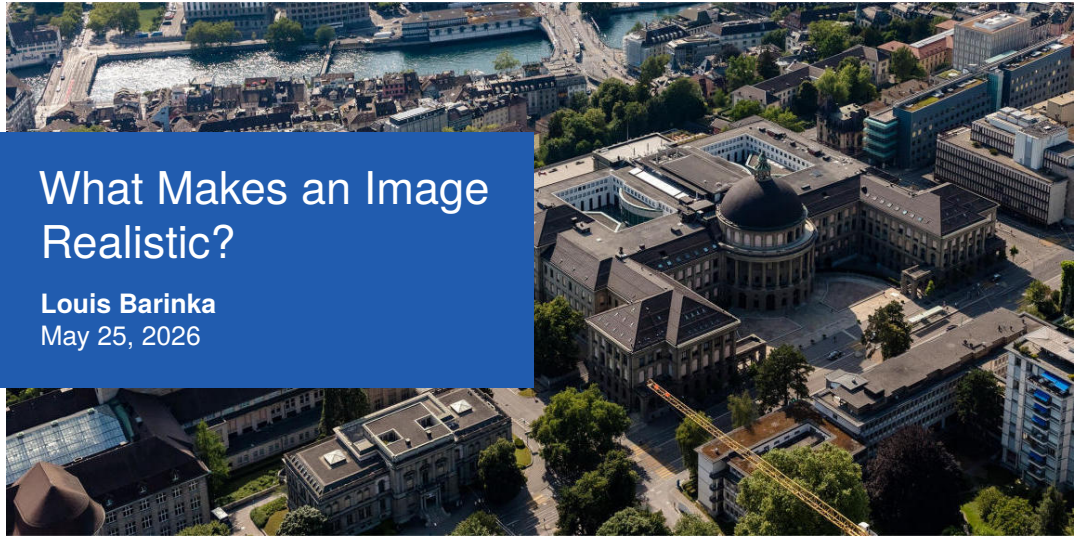
ETH zürich

Seminar presentation
after Theis (2024)



What Makes an Image Realistic?

Louis Barinka

May 25, 2026

Outline

1. Setting the stage
2. Attempt 1 — Probability
3. Attempt 2 — Weak typicality
4. Attempt 3 — Divergences
5. A universal critic

Outline

1. Setting the stage
2. Attempt 1 — Probability
3. Attempt 2 — Weak typicality
4. Attempt 3 — Divergences
5. A universal critic

The goal: a function $U(x)$

- We want $U : \mathcal{X} \to \mathbb{R}$ that scores how **realistic** an input is.

The goal: a function $U(x)$

- We want $U : \mathcal{X} \to \mathbb{R}$ that scores how **realistic** an input is.
- "Realistic" is not absolute — it lives *w.r.t. a referential* P.

The goal: a function $U(x)$

- We want $U : \mathcal{X} \to \mathbb{R}$ that scores how **realistic** an input is.
- "Realistic" is not absolute — it lives *w.r.t. a referential* P.
  - P = distribution of natural images
  - P = distribution of MNIST digits
  - P = distribution of *your* dataset

The goal: a function $U(x)$

- We want $U : \mathcal{X} \to \mathbb{R}$ that scores how **realistic** an input is.
- "Realistic" is not absolute — it lives *w.r.t. a referential* P.
  - P = distribution of natural images
  - P = distribution of MNIST digits
  - P = distribution of *your* dataset
- **Working assumption:** we have access to P.

The goal: a function $U(x)$

- We want $U : \mathcal{X} \to \mathbb{R}$ that scores how **realistic** an input is.
- "Realistic" is not absolute — it lives *w.r.t. a referential* P.
  - P = distribution of natural images
  - P = distribution of MNIST digits
  - P = distribution of *your* dataset
- **Working assumption:** we have access to P.

Question for you

If I handed you a datapoint x, how would you check that it was sampled from P?

Outline

1. Setting the stage
2. Attempt 1 — Probability
3. Attempt 2 — Weak typicality
4. Attempt 3 — Divergences
5. A universal critic

Idea: just use $P(x)$

$P(x)$ large $\Rightarrow$ keep it.

$P(x)$ small $\Rightarrow$ discard it.

Idea: just use $P(x)$

$P(x)$ large $\Rightarrow$ keep it.

$P(x)$ small $\Rightarrow$ discard it.

Very natural — this is the textbook definition of an *anomaly*.

Example 1 — a (nearly) fair coin

A bit generator emits $x = (x_1, \ldots, x_{32})$ i.i.d. with

$$P(x_n = 1) = 0.5 + \varepsilon, \qquad \varepsilon > 0.$$

Example 1 — a (nearly) fair coin

A bit generator emits $x = (x_1, \ldots, x_{32})$ i.i.d. with

$$P(x_n = 1) = 0.5 + \varepsilon, \qquad \varepsilon > 0.$$

You're debugging. Which sequence do you flag as a bug?

????????????????????????????????

????????????????????????????????

Example 1 — a (nearly) fair coin

A bit generator emits $x = (x_1, \ldots, x_{32})$ i.i.d. with

$$P(x_n = 1) = 0.5 + \varepsilon, \qquad \varepsilon > 0.$$

You're debugging. Which sequence do you flag as a bug?

11111111111111111111111111111111

????????????????????????????????

Example 1 — a (nearly) fair coin

A bit generator emits $x = (x_1, \ldots, x_{32})$ i.i.d. with

$$P(x_n = 1) = 0.5 + \varepsilon, \qquad \varepsilon > 0.$$

You're debugging. Which sequence do you flag as a bug?

`11111111111111111111111111111111`

`00010000010110010100001100011111`

Example 1 — a (nearly) fair coin

A bit generator emits $x = (x_1, \ldots, x_{32})$ i.i.d. with

$$P(x_n = 1) = 0.5 + \varepsilon, \qquad \varepsilon > 0.$$

You're debugging. Which sequence do you flag as a bug?

11111111111111111111111111111111

00010000010110010100001100011111

Example 1 — a (nearly) fair coin

A bit generator emits $x = (x_1, \ldots, x_{32})$ i.i.d. with

$$P(x_n = 1) = 0.5 + \varepsilon, \qquad \varepsilon > 0.$$

You're debugging. Which sequence do you flag as a bug?

11111111111111111111111111111111

00010000010110010100001100011111

$\arg\max_x P(x) \neq$ realistic.

“Just test more features”?

- Sure — count $k = \sum_n x_n$, check its probability.
  Catches the all-ones bug.

“Just test more features”?

- Sure — count $k = \sum_n x_n$, check its probability.
  Catches the all-ones bug.
- But every sequence has probability $\approx 2^{-N}$.

“Just test more features”?

- Sure — count $k = \sum_n x_n$, check its probability.
  Catches the all-ones bug.
- But every sequence has probability $\approx 2^{-N}$.
- If we keep adding feature tests with confidence threshold 2^{-N}…

“Just test more features”?

- Sure — count $k = \sum_n x_n$, check its probability.
  Catches the all-ones bug.
- But every sequence has probability $\approx 2^{-N}$.
- If we keep adding feature tests with confidence threshold 2^{-N}…
- …we end up rejecting *every* sequence.

“Just test more features”?

- Sure — count $k = \sum_n x_n$, check its probability.
  Catches the all-ones bug.
- But every sequence has probability $\approx 2^{-N}$.
- If we keep adding feature tests with confidence threshold 2^{-N} …
- …we end up rejecting *every* sequence.

High probability is neither necessary nor sufficient for realism.

Outline

1. Setting the stage
2. Attempt 1 — Probability
3. Attempt 2 — Weak typicality
4. Attempt 3 — Divergences
5. A universal critic

Surprise and entropy

For a discrete $x \sim P$, define the *surprise*

$$s(x) = -\log P(x).$$

Surprise and entropy

For a discrete $x \sim P$, define the *surprise*

$$s(x) = -\log P(x).$$

Rare event $\Rightarrow$ huge surprise.
Almost-certain event $\Rightarrow$ tiny surprise.

Surprise and entropy

For a discrete $x \sim P$, define the *surprise*

$$s(x) = -\log P(x).$$

Rare event $\Rightarrow$ huge surprise.
Almost-certain event $\Rightarrow$ tiny surprise.

The expected surprise is the *entropy*:

$$H[x] = \mathbb{E}_P[-\log P(x)].$$

The Asymptotic Equipartition Property

For $x_1, \ldots, x_N$ i.i.d. from P,

$$\mathbb{P}\left(\lim_{N \to \infty} -\tfrac{1}{N} \log P(x_1, \ldots, x_N) = H[x] \right) = 1.$$

The Asymptotic Equipartition Property

For $x_1, \ldots, x_N$ i.i.d. from P,

$$\mathbb{P}\left(\lim_{N \to \infty} -\tfrac{1}{N} \log P(x_1, \ldots, x_N) = H[x] \right) = 1.$$

Empirical entropy converges to the true entropy almost surely.

The typical set

$$A_\delta^N \;=\; \Big\{ x^N \;:\; \Big| -\tfrac{1}{N} \log P(x^N) - H[x] \Big| < \delta \Big\}.$$

The typical set

$$A_\delta^N \;=\; \Big\{ x^N \;:\; \Big| -\tfrac{1}{N} \log P(x^N) - H[x] \Big| < \delta \Big\}.$$

- By AEP: $\mathbb{P}(x^N \in A_\delta^N) \to 1$ for any $\delta > 0$.

The typical set

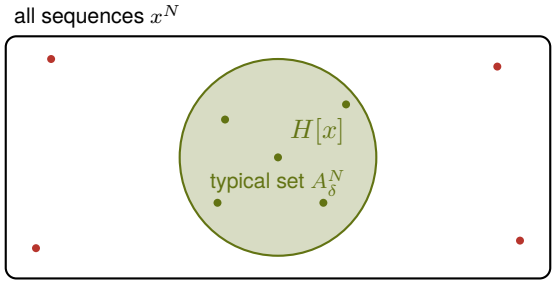
$$A_\delta^N \;=\; \Big\{ x^N \;:\; \Big| -\tfrac{1}{N}\log P(x^N) - H[x] \Big| < \delta \Big\}.$$

- By AEP: $\mathbb{P}(x^N \in A_\delta^N) \to 1$ for any $\delta > 0$.
- If $x^N \in A_\delta^N$, we call it *weakly typical*.

The typical set

$$A_\delta^N = \left\{ x^N : \left| -\tfrac{1}{N} \log P(x^N) - H[x] \right| < \delta \right\}.$$

- By AEP: $\mathbb{P}(x^N \in A_\delta^N) \to 1$ for any $\delta > 0$.
- If $x^N \in A_\delta^N$, we call it *weakly typical*.



Example 2 — typicality is not enough

Consider any distribution where every sequence x^N satisfies

$$-\log_2 P(x^N) = N.$$

Example 2 — typicality is not enough

Consider any distribution where every sequence x^N satisfies

$$-\log_2 P(x^N) = N.$$

Then *every* sequence is weakly typical.

Example 2 — typicality is not enough

Consider any distribution where every sequence x^N satisfies

$$-\log_2 P(x^N) = N.$$

Then *every* sequence is weakly typical.
In particular, our old friend

1111111111111111111111111111111

Example 2 — typicality is not enough

Consider any distribution where every sequence x^N satisfies

$$-\log_2 P(x^N) = N.$$

Then *every* sequence is weakly typical.
In particular, our old friend

1111111111111111111111111111111

passes the typicality test.

Example 2 — typicality is not enough

Consider any distribution where every sequence x^N satisfies

$$-\log_2 P(x^N) = N.$$

Then *every* sequence is weakly typical.
In particular, our old friend

`11111111111111111111111111111111`

passes the typicality test.

Take-away

Weak typicality is a *necessary* condition (for enough samples), but it is far from *sufficient* for realism.

Outline

1. Setting the stage
2. Attempt 1 — Probability
3. Attempt 2 — Weak typicality
4. Attempt 3 — Divergences
5. A universal critic

Comparing distributions instead of points

Until now we tried to judge a single x. Let's judge a *distribution* Q of candidates.

Comparing distributions instead of points

Until now we tried to judge a single x. Let's judge a *distribution* Q of candidates.

An f-divergence with convex f, $f(1) = 0$:

$$D_f[Q \,\|\, P] \;=\; \int P(x)\, f\!\left(\frac{Q(x)}{P(x)}\right) dx \;=\; \mathbb{E}_P\left[f\!\left(\frac{Q(x)}{P(x)}\right)\right].$$

Comparing distributions instead of points

Until now we tried to judge a single x. Let's judge a *distribution* Q of candidates.

An f-divergence with convex f, $f(1) = 0$:

$$D_f[Q \,\|\, P] \;=\; \int P(x)\, f\!\left(\frac{Q(x)}{P(x)}\right) dx \;=\; \mathbb{E}_P\left[f\!\left(\frac{Q(x)}{P(x)}\right)\right].$$

- $Q = P \;\Rightarrow\; D_f = 0.$

Comparing distributions instead of points

Until now we tried to judge a single x. Let's judge a *distribution* Q of candidates.

An f-divergence with convex f, $f(1) = 0$:

$$D_f[Q \,\|\, P] \;=\; \int P(x)\, f\left(\frac{Q(x)}{P(x)}\right) dx \;=\; \mathbb{E}_P\left[f\left(\frac{Q(x)}{P(x)}\right)\right].$$

- $Q = P \;\Rightarrow\; D_f = 0$.
- Jensen: $D_f[Q\|P] \geq f\left(\mathbb{E}_P\left[\frac{Q}{P}\right]\right) = f(1) = 0$.

Comparing distributions instead of points

Until now we tried to judge a single x. Let's judge a *distribution* Q of candidates.

An f-divergence with convex f, $f(1) = 0$:

$$D_f[Q \,\|\, P] \;=\; \int P(x)\, f\!\left(\frac{Q(x)}{P(x)}\right) dx \;=\; \mathbb{E}_P\left[f\!\left(\frac{Q(x)}{P(x)}\right)\right].$$

- $Q = P \;\Rightarrow\; D_f = 0$.
- Jensen: $D_f[Q\|P] \geq f\big(\mathbb{E}_P\big[\frac{Q}{P}\big]\big) = f(1) = 0$.
- $D_f = 0 \;\Rightarrow\; Q = P$, provided f attains its minimum *only* at 1.

Comparing distributions instead of points

Until now we tried to judge a single x. Let's judge a *distribution* Q of candidates.

An f-divergence with convex f, $f(1) = 0$:

$$D_f[Q \,\|\, P] \;=\; \int P(x)\, f\left(\frac{Q(x)}{P(x)}\right) dx \;=\; \mathbb{E}_P\left[f\left(\frac{Q(x)}{P(x)}\right)\right].$$

- $Q = P \;\Rightarrow\; D_f = 0$.
- Jensen: $D_f[Q\|P] \geq f\left(\mathbb{E}_P\left[\frac{Q}{P}\right]\right) = f(1) = 0$.
- $D_f = 0 \;\Rightarrow\; Q = P$, provided f attains its minimum *only* at 1.
- Generalise notions like KL-Divergence, JS-Divergence, ...

The variational view

By Fenchel duality:

$$D_f(Q \,\|\, P) \;\geq\; \mathbb{E}_Q[T(x)] - \mathbb{E}_P[f^\star(T(x))].$$

The variational view

By Fenchel duality:

$$D_f(Q \,\|\, P) \;\geq\; \mathbb{E}_Q[T(x)] - \mathbb{E}_P[f^\star(T(x))].$$

- T is any test function — a **neural network**, for instance.

The variational view

By Fenchel duality:

$$D_f(Q \,\|\, P) \;\geq\; \mathbb{E}_Q[T(x)] - \mathbb{E}_P[f^\star(T(x))].$$

- T is any test function — a **neural network**, for instance.
- If we have lots of samples from Q and P, we estimate both expectations by Monte Carlo.

The variational view

By Fenchel duality:

$$D_f(Q \,\|\, P) \;\geq\; \mathbb{E}_Q[T(x)] - \mathbb{E}_P[f^\star(T(x))].$$

- T is any test function — a **neural network**, for instance.
- If we have lots of samples from Q and P, we estimate both expectations by Monte Carlo.
- Training T to maximize the bound.

Example: KL divergence

Take $f(u) = u \log u$.

$$D_{\mathrm{KL}}(Q\|P) \;=\; \mathbb{E}_Q\left[\log \frac{Q(x)}{P(x)}\right].$$

Example: KL divergence

Take $f(u) = u \log u$.

$$D_{\mathrm{KL}}(Q\|P) \;=\; \mathbb{E}_Q\left[\log \frac{Q(x)}{P(x)}\right].$$

Setting $T_Q(x) = \log Q(x) - \log P(x) + 1$ makes the variational bound **tight**.

Example: KL divergence

Take $f(u) = u \log u$.

$$D_{\mathrm{KL}}(Q\|P) \;=\; \mathbb{E}_Q\left[\log \frac{Q(x)}{P(x)}\right].$$

Setting $T_Q(x) = \log Q(x) - \log P(x) + 1$ makes the variational bound **tight**.

Catch

Divergences need *many samples* from Q. Often we have one image, or a handful. $\Rightarrow$ impractical as a per-sample critic.

Outline

1. Setting the stage
2. Attempt 1 — Probability
3. Attempt 2 — Weak typicality
4. Attempt 3 — Divergences
5. A universal critic

The proposal

Borrow tools from **algorithmic information theory**:

$$U(x) \;=\; -\log P(x) \;-\; K(x)$$

The proposal

Borrow tools from **algorithmic information theory**:

$$U(x) \;=\; -\log P(x) \;-\; K(x)$$

The proposal

Borrow tools from **algorithmic information theory**:

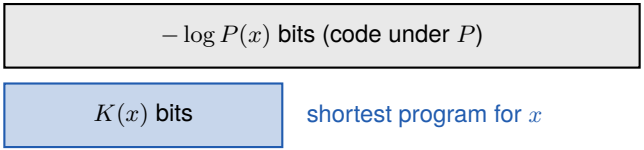
$$U(x) \;=\; -\log P(x) \;-\; K(x)$$

$-\log P(x)$ *encodes* x; $\quad K(x)$ *describes* x.

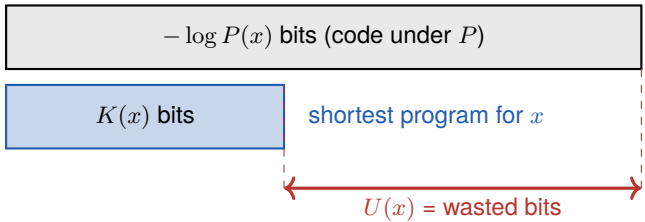
Picturing it

$-\log P(x)$ bits (code under P)

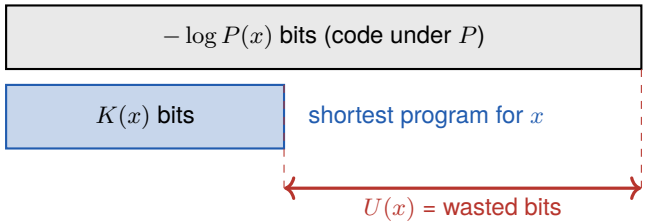
Picturing it



Picturing it



Picturing it

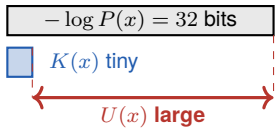


- x realistic w.r.t. $P \Rightarrow P$ compresses x about as well as the best program.
- x unrealistic $\Rightarrow P$ wastes many bits compared to the true structure of x.

Example 1, revisited

P = fair coin, $N = 32$. Both sequences satisfy $-\log_2 P(x) = 32$.

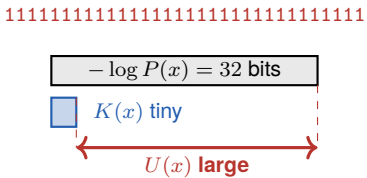
11111111111111111111111111111111

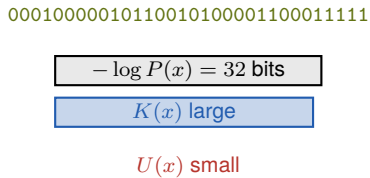


$\Rightarrow$ **unrealistic**

Example 1, revisited

P = fair coin, $N = 32$. Both sequences satisfy $-\log_2 P(x) = 32$.





A probabilistic detour: Solomonoff

Enumerate all computable discrete distributions $\{Q_n\}$ and give them a prior $\{\pi_n\}$ with $\sum_n \pi_n \leq 1$. Define

$$S(x) \;=\; \sum_n \pi_n \, Q_n(x).$$

A probabilistic detour: Solomonoff

Enumerate all computable discrete distributions $\{Q_n\}$ and give them a prior $\{\pi_n\}$ with $\sum_n \pi_n \leq 1$. Define

$$S(x) \;=\; \sum_n \pi_n \, Q_n(x).$$

A giant mixture over *all hypotheses you could ever write down*.

A probabilistic detour: Solomonoff

Enumerate all computable discrete distributions $\{Q_n\}$ and give them a prior $\{\pi_n\}$ with $\sum_n \pi_n \leq 1$. Define

$$S(x) \;=\; \sum_n \pi_n \, Q_n(x).$$

A giant mixture over *all hypotheses you could ever write down*.

Coding theorem. Under mild conditions on π_n,

$$K(x) \;=\; -\log S(x).$$

What happens when x is drawn from some Q?

Plug $K(x) = -\log S(x)$ into U, then take $\mathbb{E}_{x \sim Q}$.

What happens when x is drawn from some Q?

Plug $K(x) = -\log S(x)$ into U, then take $\mathbb{E}_{x\sim Q}$.
Upper bound (non-negativity of KL between Q and S):

$$\begin{aligned}\mathbb{E}_Q[U(x)] &= \mathbb{E}_Q[\log S(x) - \log P(x)] \\ &\leq \mathbb{E}_Q[\log Q(x) - \log P(x)] = D_{\mathrm{KL}}(Q \,\|\, P).\end{aligned}$$

What happens when x is drawn from some Q?

Plug $K(x) = -\log S(x)$ into U, then take $\mathbb{E}_{x\sim Q}$.
Upper bound (non-negativity of KL between Q and S):

$$\begin{aligned}\mathbb{E}_Q[U(x)] &= \mathbb{E}_Q[\log S(x) - \log P(x)] \\ &\leq \mathbb{E}_Q[\log Q(x) - \log P(x)] = D_{\mathrm{KL}}(Q \,\|\, P).\end{aligned}$$

Lower bound (keep only the Q-term in the mixture S):

$$\begin{aligned}\mathbb{E}_Q[U(x)] &\geq \mathbb{E}_Q[\log(\pi_Q Q(x)) - \log P(x)] \\ &= D_{\mathrm{KL}}(Q \,\|\, P) - \log \tfrac{1}{\pi_Q}.\end{aligned}$$

The sandwich

$$\boxed{D_{\mathrm{KL}}(Q\|P) - \log \tfrac{1}{\pi_Q} \;\leq\; \mathbb{E}_{x\sim Q}[U(x)] \;\leq\; D_{\mathrm{KL}}(Q\|P).}$$

The sandwich

$$\boxed{D_{\mathrm{KL}}(Q\|P) - \log \tfrac{1}{\pi_Q} \;\le\; \mathbb{E}_{x\sim Q}[U(x)] \;\le\; D_{\mathrm{KL}}(Q\|P).}$$

Choose the prior

$$\pi_Q \;=\; \sum_{p\,:\,Q_p = Q} 2^{-|p|} \quad\Longrightarrow\quad -\log_2 \pi_Q \;\le\; K(Q).$$

The sandwich

$$\boxed{D_{\mathrm{KL}}(Q\|P) - \log \tfrac{1}{\pi_Q} \;\le\; \mathbb{E}_{x\sim Q}[U(x)] \;\le\; D_{\mathrm{KL}}(Q\|P).}$$

Choose the prior

$$\pi_Q \;=\; \sum_{p\,:\,Q_p = Q} 2^{-|p|} \quad\Longrightarrow\quad -\log_2 \pi_Q \;\le\; K(Q).$$

Then:

$$\boxed{D_{\mathrm{KL}}(Q\|P) - K(Q) \;\le\; \mathbb{E}_{x\sim Q}[U(x)] \;\le\; D_{\mathrm{KL}}(Q\|P).}$$

Many samples: a batched critic

What if we observe B i.i.d. draws $x^B = (x_1, \ldots, x_B)$ from Q?

Many samples: a batched critic

What if we observe B i.i.d. draws $x^B = (x_1, \ldots, x_B)$ from Q?

$$U^B(x^B) \;=\; \log \sum_n \pi_n \prod_b Q_n(x_b) \;-\; \log \prod_b P(x_b).$$

Many samples: a batched critic

What if we observe B i.i.d. draws $x^B = (x_1, \ldots, x_B)$ from Q?

$$U^B(x^B) \;=\; \log \sum_n \pi_n \prod_b Q_n(x_b) \;-\; \log \prod_b P(x_b).$$

Sandwich, batched:

$$\boxed{D_{\mathrm{KL}}(Q\|P) \;-\; \tfrac{K(Q)}{B} \;\le\; \tfrac{1}{B}\,\mathbb{E}_{Q^B}\left[U^B(x^B)\right] \;\le\; D_{\mathrm{KL}}(Q\|P).}$$

Takeaways

- ✖ $P(x)$ alone: rewards the *most probable*, not the realistic.

Takeaways

- ✖ $P(x)$ alone: rewards the *most probable*, not the realistic.
- ✖ Weak typicality: *necessary*, not sufficient.

Takeaways

- ✖ $P(x)$ alone: rewards the *most probable*, not the realistic.
- ✖ Weak typicality: *necessary*, not sufficient.
- ✖ D_f-divergences: aggregate, need many samples.

Takeaways

- ✗ $P(x)$ alone: rewards the *most probable*, not the realistic.
- ✗ Weak typicality: *necessary*, not sufficient.
- ✗ D_f-divergences: aggregate, need many samples.
- ✓ Universal critic on *one* sample: excellent when $K(Q)$ is *low*.

Takeaways

- ✗ $P(x)$ alone: rewards the *most probable*, not the realistic.
- ✗ Weak typicality: *necessary*, not sufficient.
- ✗ D_f-divergences: aggregate, need many samples.
- ✓ Universal critic on *one* sample: excellent when $K(Q)$ is *low*.
- ✓ Higher $K(Q) \Rightarrow$ need more samples (batched critic U^B).

Takeaways

- ✖ $P(x)$ alone: rewards the *most probable*, not the realistic.
- ✖ Weak typicality: *necessary*, not sufficient.
- ✖ D_f-divergences: aggregate, need many samples.
- ✔ Universal critic on *one* sample: excellent when $K(Q)$ is *low*.
- ✔ Higher $K(Q) \Rightarrow$ need more samples (batched critic U^B).
- ★ $K(x)$ is *not computable*, but U is our **north star** for designing realism criteria.