Compression and Contraction

Adway Girish
Information Theory Lab

EPFL Information Processing Group

September 9, 2024
IPG PhD Review

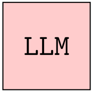
Outline

1. Prompt compression for black-box language models
2. Input-entropy-constrained capacity
3. Joint range of divergences
4. Closing remarks

Outline

1. Prompt compression for black-box language models
2. Input-entropy-constrained capacity
3. Joint range of divergences
4. Closing remarks

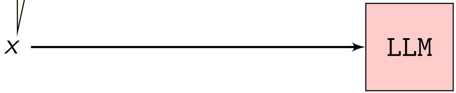
Prompt compression



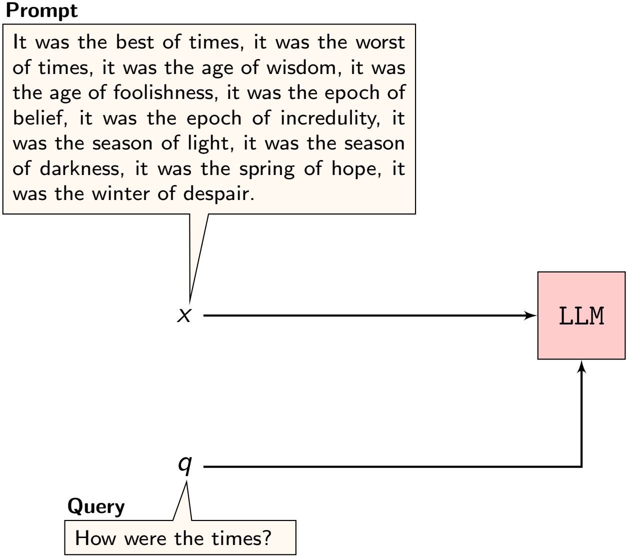
Prompt compression

Prompt

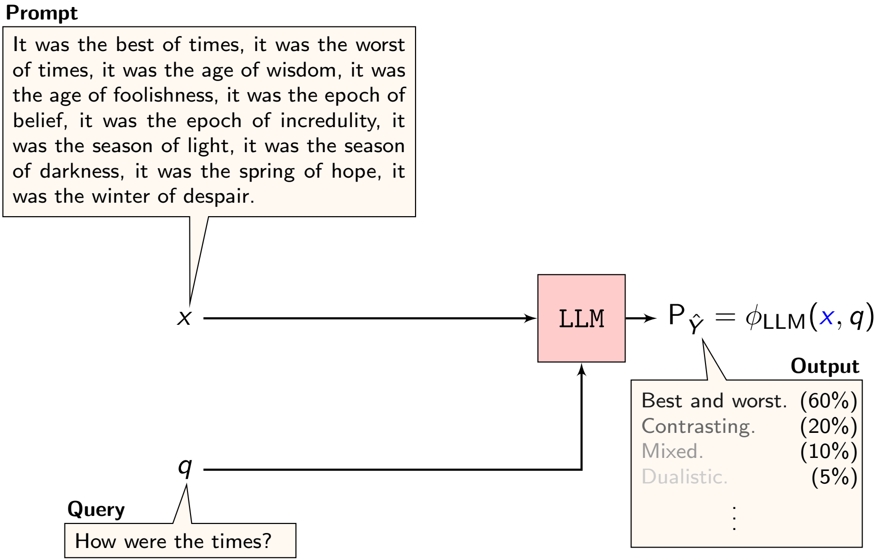
It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of light, it was the season of darkness, it was the spring of hope, it was the winter of despair.



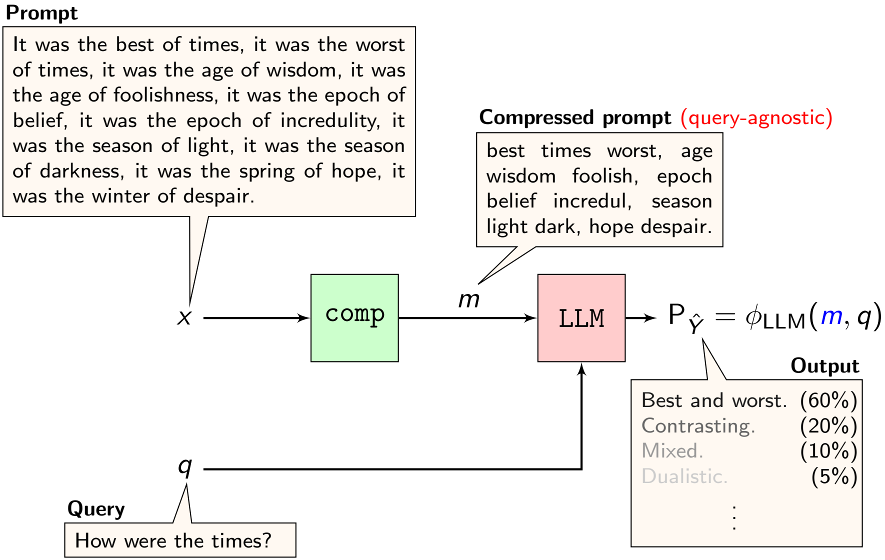
Prompt compression



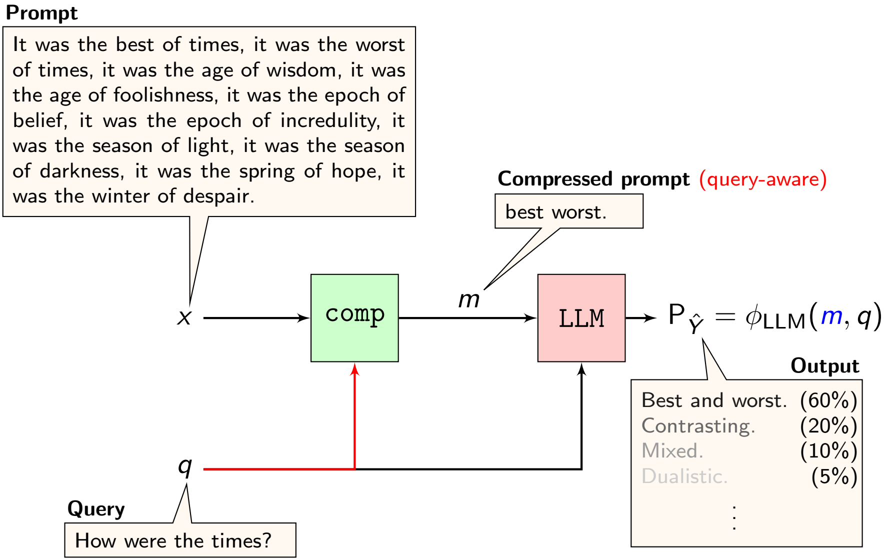
Prompt compression



Prompt compression: query-agnostic



Prompt compression: query-aware

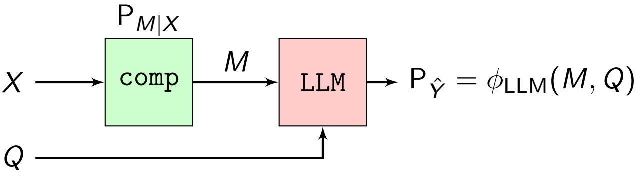


Prompt compression: rate-distortion formulation

- $(X, Q, Y) \sim \mathrm{P}_{XQY} = \mathrm{P}_{XQ}\,\mathrm{P}_{Y|XQ}$ $\qquad Y =$ “true answer”

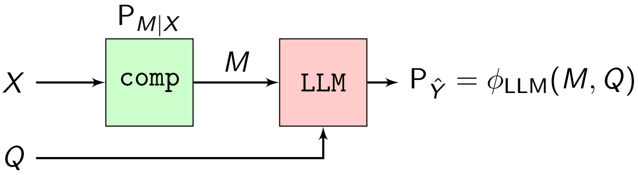
Prompt compression: rate-distortion formulation

- $(X, Q, Y) \sim \mathsf{P}_{XQY} = \mathsf{P}_{XQ}\,\mathsf{P}_{Y|XQ}$ $\quad$ $Y =$ "true answer"



Prompt compression: rate-distortion formulation

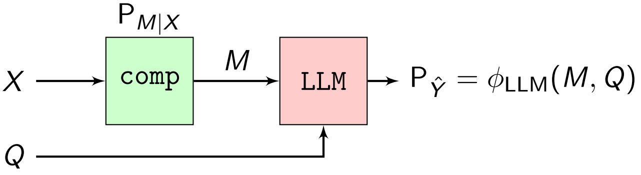
- $(X, Q, Y) \sim \mathsf{P}_{XQY} = \mathsf{P}_{XQ}\,\mathsf{P}_{Y|XQ}$ $\qquad Y =$ "true answer"



- Compression with side-information

Prompt compression: rate-distortion formulation

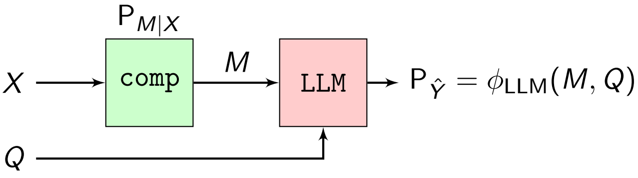
- $(X, Q, Y) \sim \mathrm{P}_{XQY} = \mathrm{P}_{XQ}\,\mathrm{P}_{Y|XQ}$ $\qquad Y =$ "true answer"



- Compression with side-information
  for a fixed decoder, "$(m, q) \mapsto \phi_{\mathsf{LLM}}(m, q)$"

Prompt compression: rate-distortion formulation

- $(X, Q, Y) \sim \mathrm{P}_{XQY} = \mathrm{P}_{XQ}\,\mathrm{P}_{Y|XQ}$ $\qquad Y =$ "true answer"

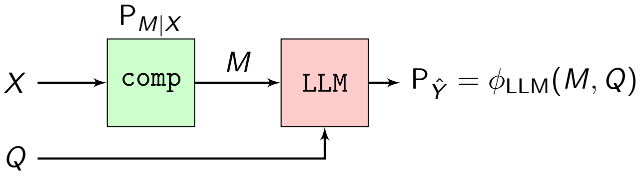


- Compression with side-information
  for a fixed decoder, "$(m, q) \mapsto \phi_{\mathrm{LLM}}(m, q)$"

- Performance metrics:

Prompt compression: rate-distortion formulation

- $(X, Q, Y) \sim \mathrm{P}_{XQY} = \mathrm{P}_{XQ}\,\mathrm{P}_{Y|XQ}$ $\qquad$ $Y =$ "true answer"



- Compression with side-information

  for a fixed decoder, "$(m, q) \mapsto \phi_{\mathsf{LLM}}(m, q)$"

- Performance metrics:

$$\text{rate} = \mathbb{E}\left[\frac{\mathrm{len}(M)}{\mathrm{len}(X)}\right] \qquad \text{distortion} = \mathbb{E}\left[\mathrm{d}\big(Y, \phi_{\mathsf{LLM}}(M, Q)\big)\right]$$

Distortion-rate function

- $$\text{rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{distortion} = \mathbb{E}\left[\text{d}(Y, \phi_{\text{LLM}}(M, Q))\right]$$

Distortion-rate function

- $$\text{rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{distortion} = \mathbb{E}\left[\text{d}(Y, \phi_{\text{LLM}}(M, Q))\right]$$

- $$\begin{aligned} D^*(R) = \quad & \inf_{\text{P}_{M|X}} && \mathbb{E}\left[\text{d}\big(Y, \phi_{\text{LLM}}(M, Q)\big)\right] \\ & \text{s.t.} && \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \le R, \text{ and} \\ & && \text{P}_{M|X} \text{ ``is a compressor''} \end{aligned}$$

Distortion-rate function

- $$\text{rate} = \mathbb{E}\left[\frac{\mathrm{len}(M)}{\mathrm{len}(X)}\right] \qquad \text{distortion} = \mathbb{E}\left[\mathrm{d}(Y, \phi_{\mathrm{LLM}}(M, Q))\right]$$
- $$\begin{aligned} D^*(R) = \quad & \inf_{\mathrm{P}_{M|X}} \quad \mathbb{E}\left[\mathrm{d}\big(Y, \phi_{\mathrm{LLM}}(M, Q)\big)\right] \\ & \text{s.t.} \quad \mathbb{E}\left[\frac{\mathrm{len}(M)}{\mathrm{len}(X)}\right] \le R, \text{ and} \\ & \qquad\ \mathrm{P}_{M|X} \text{ "is a compressor"} \end{aligned}$$
- Linear program, but large dimension

Distortion-rate function

- $$\text{rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{distortion} = \mathbb{E}\left[\text{d}(Y, \phi_{\text{LLM}}(M, Q))\right]$$

- $$\begin{aligned} D^*(R) = \quad \inf_{\text{P}_{M|X}} \quad & \mathbb{E}\left[\text{d}\big(Y, \phi_{\text{LLM}}(M, Q)\big)\right] \\ \text{s.t.} \quad & \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \leq R, \text{ and} \\ & \text{P}_{M|X} \text{ ``is a compressor''} \end{aligned}$$

- Linear program, but large dimension $\approx 32{,}000^{10}$

Distortion-rate function

- $$\text{rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{distortion} = \mathbb{E}\left[\text{d}(Y, \phi_{\text{LLM}}(M, Q))\right]$$
- $$\begin{aligned} D^*(R) = \quad & \inf_{\mathsf{P}_{M|X}} \quad \mathbb{E}\left[\text{d}\big(Y, \phi_{\text{LLM}}(M, Q)\big)\right] \\ & \text{s.t.} \quad \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \leq R, \text{ and} \\ & \qquad\; \mathsf{P}_{M|X} \text{ ``is a compressor''} \end{aligned}$$
- Linear program, but large dimension $\approx 32{,}000^{10}$
- Dual:
$$D^*(R) \; = \; \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \min_{m \in \mathcal{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \, \boldsymbol{R}_{x,m} \right] \right\}$$

Distortion-rate function

- $$\text{rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{distortion} = \mathbb{E}\left[\text{d}(Y, \phi_{\text{LLM}}(M, Q))\right]$$

- $$\begin{aligned} D^*(R) = \quad \inf_{\mathsf{P}_{M|X}} \quad & \mathbb{E}\left[\text{d}\big(Y, \phi_{\text{LLM}}(M, Q)\big)\right] \\ \text{s.t.} \quad & \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \le R, \text{ and} \\ & \mathsf{P}_{M|X} \text{ ``is a compressor''} \end{aligned}$$

- Linear program, but large dimension $\approx 32{,}000^{10}$

- Dual:
$$D^*(R) \;=\; \sup_{\lambda \ge 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \min_{m \in \overset{\text{all possible ``compressions'' of } x}{\overset{\downarrow}{\mathcal{M}_x}}} \left[\boldsymbol{D}_{x,m} + \lambda \, \boldsymbol{R}_{x,m} \right] \right\}$$

Distortion-rate function

- $$\text{rate} = \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \qquad \text{distortion} = \mathbb{E}\left[\text{d}(Y, \phi_{\text{LLM}}(M, Q))\right]$$

- $$\begin{aligned} D^*(R) = \quad & \inf_{\text{P}_{M|X}} \quad \mathbb{E}\left[\text{d}\big(Y, \phi_{\text{LLM}}(M, Q)\big)\right] \\ & \text{s.t.} \quad \mathbb{E}\left[\frac{\text{len}(M)}{\text{len}(X)}\right] \leq R, \text{ and} \\ & \qquad\; \text{P}_{M|X} \text{ ``is a compressor''} \end{aligned}$$

- Linear program, but large dimension $\approx 32{,}000^{10}$

- Dual:

$$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \min_{m \in \overset{\text{all possible ``compressions'' of } x}{\mathcal{M}_x}} \left[ \underbrace{\boldsymbol{D}_{x,m}}_{\text{``normalized'' distortion}} + \lambda \underbrace{\boldsymbol{R}_{x,m}}_{\text{rate}} \right] \right\}$$

"normalized" distortion, rate on compressing $x \mapsto m$

Distortion-rate function: geometric solution via dual

- Dual: $$D^*(R) \;=\; \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \min_{m \in \mathcal{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right] \right\}$$

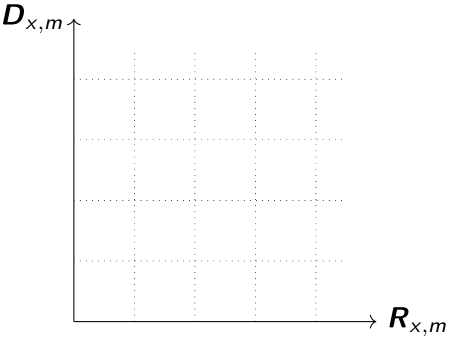
Distortion-rate function: geometric solution via dual

- Dual: $$D^*(R) \;=\; \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \underbrace{\min_{m \in \mathcal{M}_x} [\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m}]}_{\text{for fixed } (\lambda, x)} \right\}$$

Distortion-rate function: geometric solution via dual

- Dual: $$D^*(R) \;=\; \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \underbrace{\min_{m \in \mathcal{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right]}_{\text{for fixed } (\lambda, x)} \right\}$$

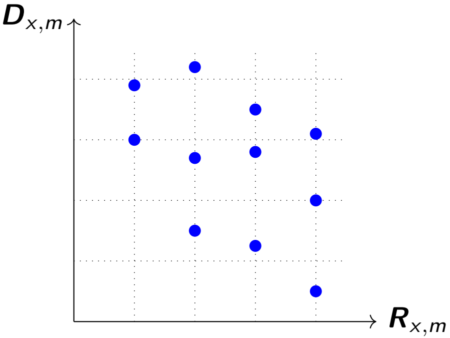
- Fix $\lambda \geq 0$, $x \in \mathcal{X}$



Distortion-rate function: geometric solution via dual

- Dual: $$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \underbrace{\min_{m \in \mathcal{M}_x} [\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m}]}_{\text{for fixed } (\lambda, x)} \right\}$$

- Fix $\lambda \geq 0$, $x \in \mathcal{X}$

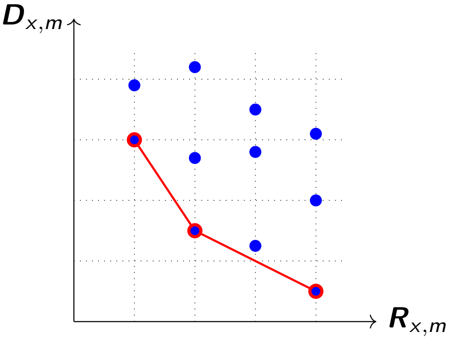


- Relevant points: $32{,}000^{10}$

Distortion-rate function: geometric solution via dual

- Dual: $$D^*(R) = \sup_{\lambda \geq 0} \Big\{ -\lambda R + \sum_{x \in \mathcal{X}} \underbrace{\min_{m \in \mathcal{M}_x} [\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m}]}_{\text{for fixed } (\lambda, x)} \Big\}$$

- Fix $\lambda \geq 0$, $x \in \mathcal{X}$

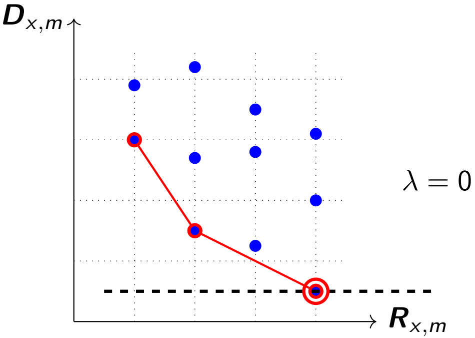


- Relevant points: $32{,}000^{10}$

Distortion-rate function: geometric solution via dual

- Dual: $$D^*(R) \;=\; \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \underbrace{\min_{m \in \mathcal{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m} \right]}_{\text{for fixed } (\lambda, x)} \right\}$$

- Fix $\lambda \geq 0$, $x \in \mathcal{X}$

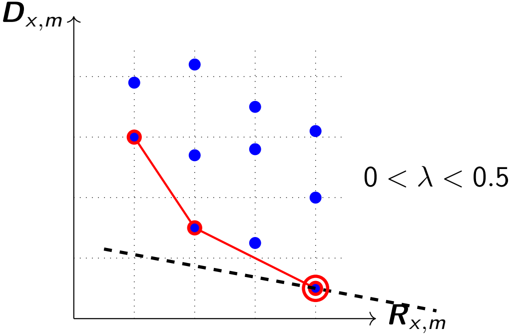


- Relevant points: $32{,}000^{10}$

Distortion-rate function: geometric solution via dual

- Dual: $$D^*(R) \;=\; \sup_{\lambda \geq 0} \Bigg\{ -\lambda R + \sum_{x \in \mathcal{X}} \underbrace{\min_{m \in \mathcal{M}_x} [\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m}]}_{\text{for fixed } (\lambda, x)} \Bigg\}$$

- Fix $\lambda \geq 0$, $x \in \mathcal{X}$

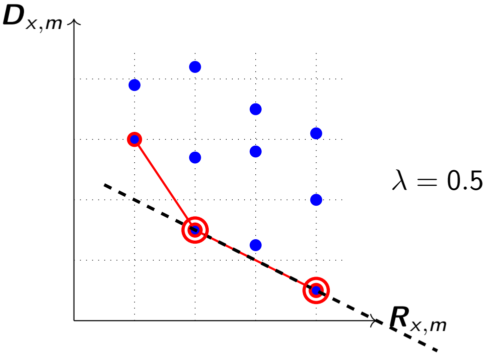


- Relevant points: $32{,}000^{10}$

Distortion-rate function: geometric solution via dual

- Dual: $$D^*(R) \;=\; \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \underbrace{\min_{m \in \mathcal{M}_x} \left[\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m}\right]}_{\text{for fixed } (\lambda, x)} \right\}$$

- Fix $\lambda \geq 0$, $x \in \mathcal{X}$

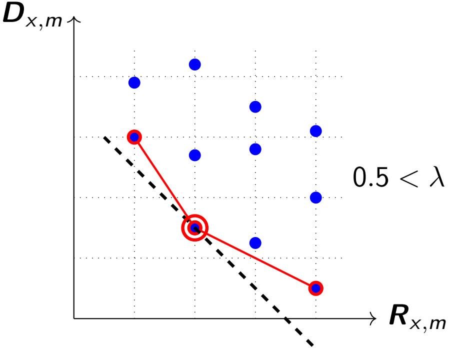


- Relevant points: $32{,}000^{10}$

Distortion-rate function: geometric solution via dual

- Dual: $$D^*(R) \;=\; \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \underbrace{\min_{m \in \mathcal{M}_x} [\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m}]}_{\text{for fixed } (\lambda, x)} \right\}$$

- Fix $\lambda \geq 0$, $x \in \mathcal{X}$

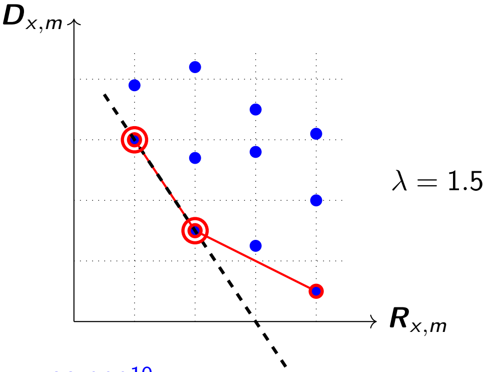


$0.5 < \lambda < 1.5$

- Relevant points: $32{,}000^{10}$

Distortion-rate function: geometric solution via dual

- Dual: $$D^*(R) = \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \underbrace{\min_{m \in \mathcal{M}_x} [\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m}]}_{\text{for fixed } (\lambda, x)} \right\}$$
- Fix $\lambda \geq 0$, $x \in \mathcal{X}$

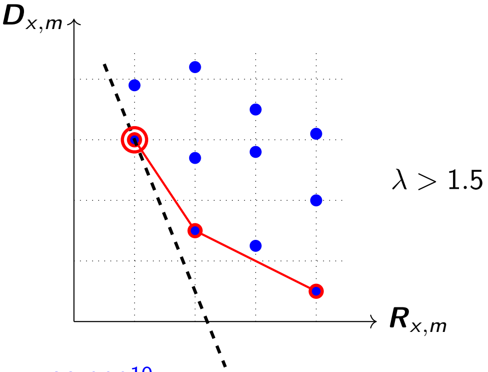


- Relevant points: $32{,}000^{10}$

Distortion-rate function: geometric solution via dual

- Dual: $$D^*(R) = \sup_{\lambda \geq 0} \Big\{ -\lambda R + \sum_{x\in\mathcal{X}} \underbrace{\min_{m\in\mathcal{M}_x} [\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m}]}_{\text{for fixed } (\lambda,x)} \Big\}$$

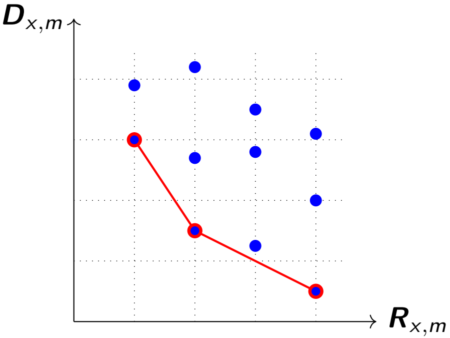
- Fix $\lambda \geq 0$, $x \in \mathcal{X}$



- Relevant points: $32{,}000^{10}$

Distortion-rate function: geometric solution via dual

- Dual: $$D^*(R) \;=\; \sup_{\lambda \geq 0} \Big\{ -\lambda R + \sum_{x \in \mathcal{X}} \underbrace{\min_{m \in \mathcal{M}_x} [\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m}]}_{\text{for fixed } (\lambda, x)} \Big\}$$
- Fix $\lambda \geq 0$, $x \in \mathcal{X}$

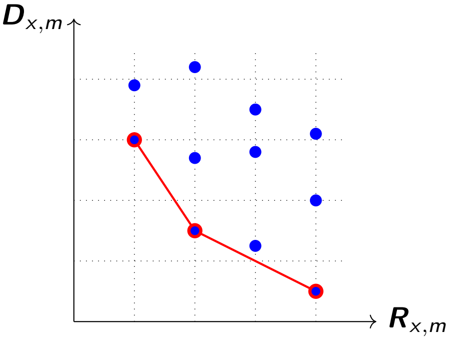


- Relevant points: $32{,}000^{10} \to 10$

Distortion-rate function: geometric solution via dual

- Dual: $$D^*(R) \;=\; \sup_{\lambda \geq 0} \Big\{ -\lambda R + \sum_{x \in \mathcal{X}} \underbrace{\min_{m \in \mathcal{M}_x} [\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m}]}_{\text{for fixed } (\lambda, x)} \Big\}$$

- Fix $\lambda \geq 0$, $x \in \mathcal{X}$

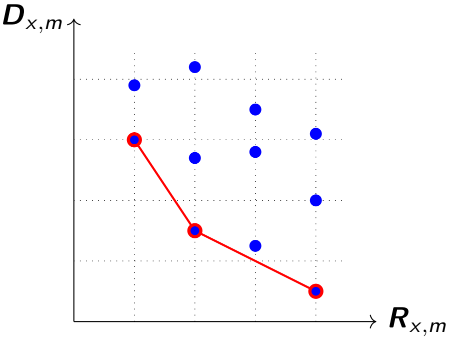


- Relevant points: $32{,}000^{10} \to 10$,
  only finitely many λ

Distortion-rate function: geometric solution via dual

- Dual: $$D^*(R) \;=\; \sup_{\lambda \geq 0} \left\{ -\lambda R + \sum_{x \in \mathcal{X}} \underbrace{\min_{m \in \mathcal{M}_x} [\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m}]}_{\text{for fixed } (\lambda, x)} \right\}$$

- Fix $\lambda \geq 0$, $x \in \mathcal{X}$

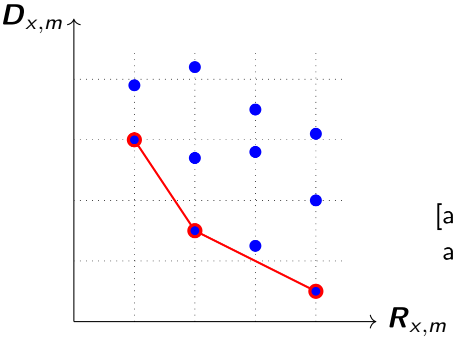


- Relevant points: $32{,}000^{10} \to 10$, only finitely many λ $\quad (2^{10} \to 10)$

Distortion-rate function: geometric solution via dual

- Dual: $$D^*(R) = \sup_{\lambda \geq 0} \Big\{ -\lambda R + \sum_{x \in \mathcal{X}} \underbrace{\min_{m \in \mathcal{M}_x} [\boldsymbol{D}_{x,m} + \lambda \boldsymbol{R}_{x,m}]}_{\text{for fixed } (\lambda, x)} \Big\}$$

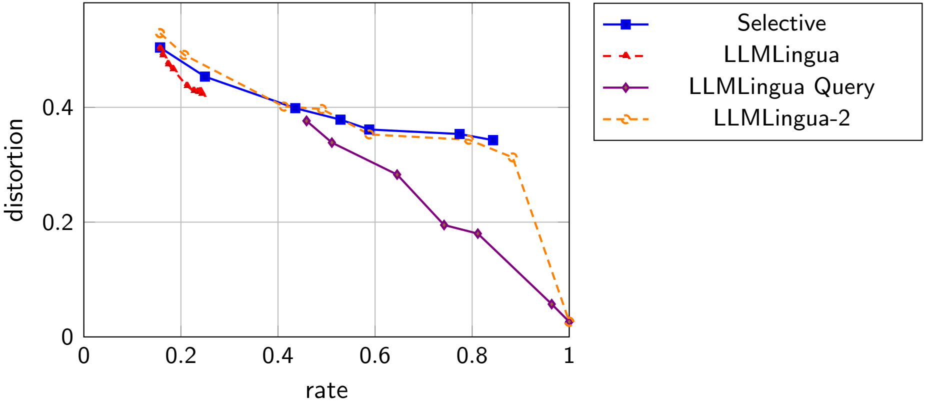
- Fix $\lambda \geq 0$, $x \in \mathcal{X}$



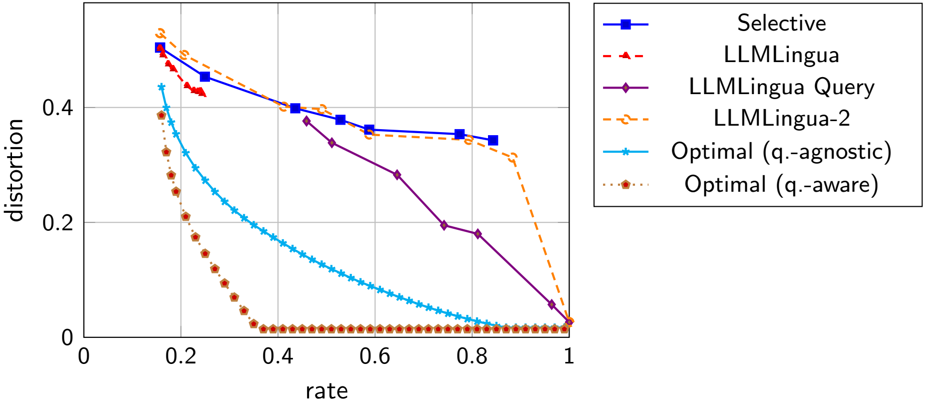
[apple $\mapsto$ app, ale, pe;
apple $\not\mapsto$ pale, red, lp]

- Relevant points: $32{,}000^{10} \to 10$, only finitely many λ $\qquad (2^{10} \to 10)$

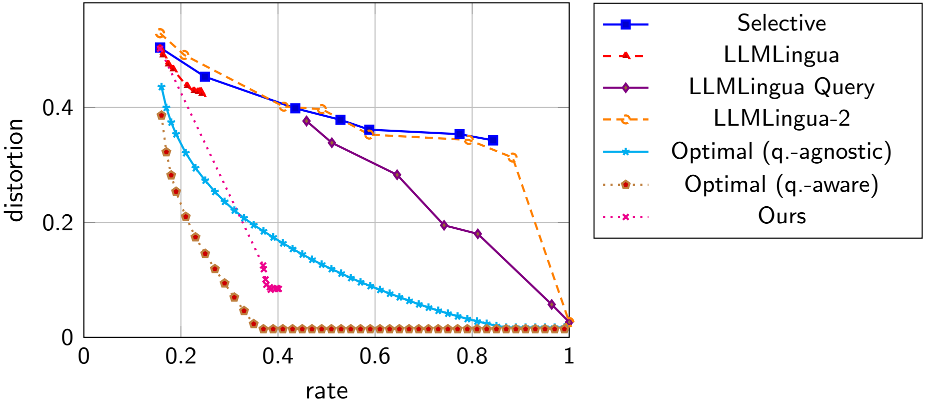
Experimental results



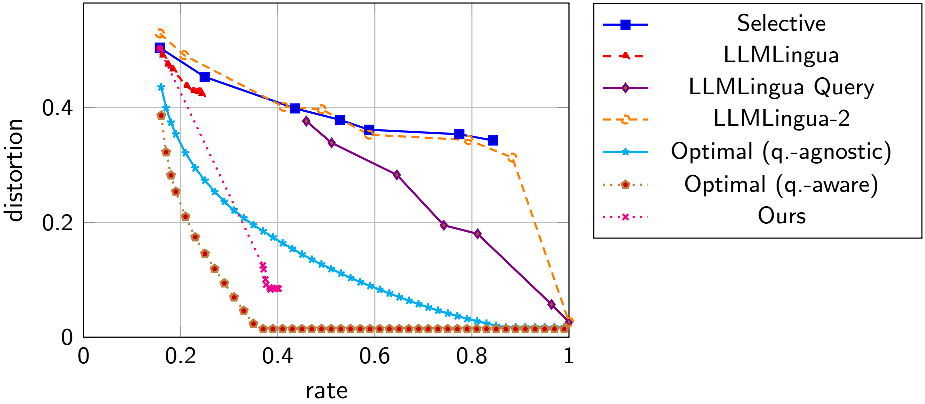
Experimental results



Experimental results



Experimental results



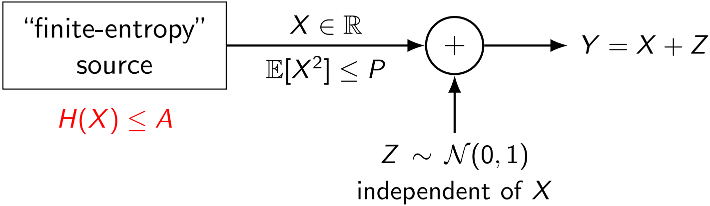
A.G.*, A.Nagle*, M.Bondaschi, M.Gastpar, A.V.Makkuva, H.Kim, "Fundamental Limits of Prompt Compression: A Rate-Distortion Framework for Black-Box Language Models."
— ICML 2024 Workshop on Theoretical Foundations of Foundation Models **[Oral]**
— under review at `[conference]`

Segue to a contraction problem

- Optimization 101. . .

Segue to a contraction problem

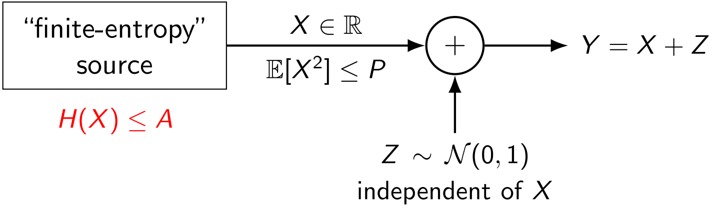
- Optimization 101…thanks to a different problem:



Outline

Input-entropy-constrained channel capacity

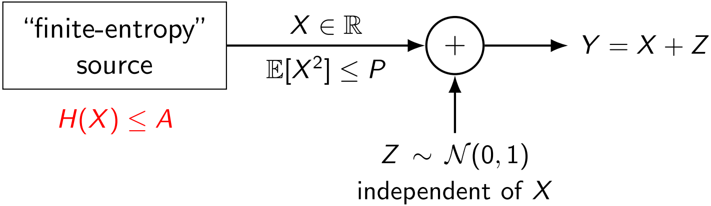
- A contraction problem in communication



$$C_H(A,P) = \sup_{\substack{\mathrm{P}_X:\ \mathbb{E}[X^2] \leq P \\ H(X) \leq A}} I(X;Y)$$

Input-entropy-constrained channel capacity

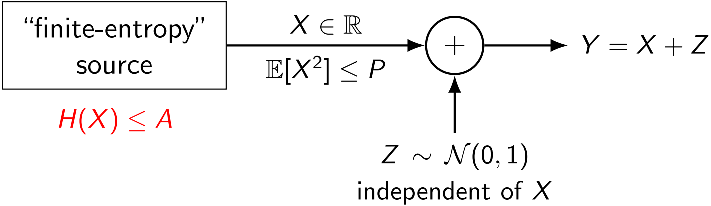
- A contraction problem in communication



$$C_H(A,P) = \sup_{\substack{P_X:\ \mathbb{E}[X^2] \leq P \\ H(X) \leq A}} I(X;Y)$$

- Cardinality bounds?

- A contraction problem in communication

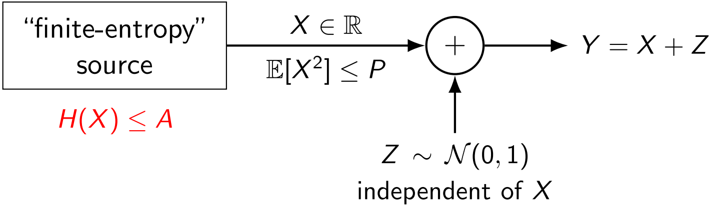


$$C_H(A,P) = \sup_{\substack{P_X:\ \mathbb{E}[X^2] \leq P \\ H(X) \leq A}} I(X;Y)$$

- Cardinality bounds? Finite support?

Input-entropy-constrained channel capacity

- A contraction problem in communication



$$C_H(A,P) = \sup_{\substack{P_X:\ \mathbb{E}[X^2] \le P \\ H(X) \le A}} I(X;Y)$$

- Cardinality bounds? Finite support?
- A nontrivial upper bound better than

$$F_I(A,P) = \sup_{\substack{P_{WX}:\ \mathbb{E}[X^2] \le P \\ I(W;X) \le A}} I(W;Y) \qquad ?$$

Aside on data processing inequalities

Fix $\mathsf{P}_{Y|X}$

Aside on data processing inequalities

Fix $\mathrm{P}_{Y|X}$

- DPI: for any P_{WX}, $I(W;Y) \leq I(W;X)$

Aside on data processing inequalities

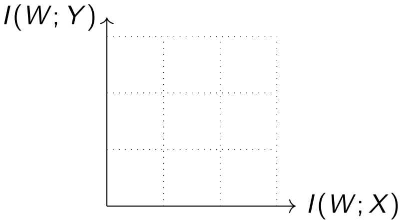
Fix $\mathsf{P}_{Y|X}$

- DPI: for any P_{WX}, $I(W;Y) \le I(W;X)$
- Data processing *function*: $F_I(t) = \sup_{\mathsf{P}_{WX}:\, I(W;X) \le t} I(W;Y)$

Aside on data processing inequalities

Fix $\mathsf{P}_{Y|X}$

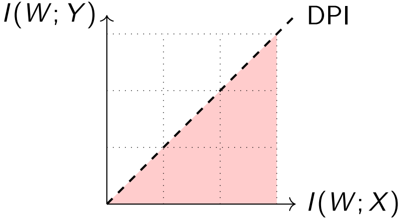
- DPI: for any P_{WX}, $I(W;Y) \le I(W;X)$
- Data processing *function*: $F_I(t) = \sup_{\mathsf{P}_{WX}:\, I(W;X) \le t} I(W;Y)$



Aside on data processing inequalities

Fix $\mathsf{P}_{Y|X}$

- DPI: for any P_{WX}, $I(W;Y) \leq I(W;X)$
- Data processing *function*: $F_I(t) = \sup_{\mathsf{P}_{WX}:\, I(W;X) \leq t} I(W;Y)$

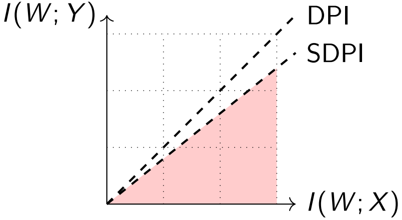


Aside on data processing inequalities

Fix $\mathrm{P}_{Y|X}$

- DPI: for any P_{WX}, $I(W;Y) \leq I(W;X)$

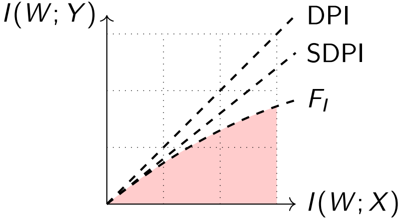
- Data processing *function*: $F_I(t) = \sup_{\mathrm{P}_{WX}:\, I(W;X) \leq t} I(W;Y)$



Aside on data processing inequalities

Fix $\mathrm{P}_{Y|X}$

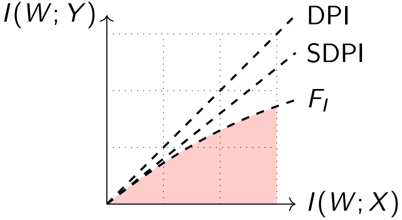
- DPI: for any P_{WX}, $I(W;Y) \leq I(W;X)$
- Data processing *function*: $F_I(t) = \sup_{\mathrm{P}_{WX}:\, I(W;X) \leq t} I(W;Y)$



Aside on data processing inequalities

Fix $\mathsf{P}_{Y|X}$

- DPI: for any P_{WX}, $I(W;Y) \leq I(W;X)$
- Data processing *function*: $F_I(t) = \sup_{\mathsf{P}_{WX}:\, I(W;X) \leq t} I(W;Y)$

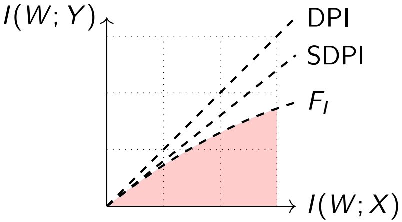


- Also DPI: for any $\mathsf{P}_X, \mathsf{Q}_X$, $D_f(\mathsf{Q}_Y \,\|\, \mathsf{P}_Y) \leq D_f(\mathsf{Q}_X \,\|\, \mathsf{P}_X)$

Aside on data processing inequalities

Fix $\mathrm{P}_{Y|X}$

- DPI: for any P_{WX}, $I(W;Y) \leq I(W;X)$
- Data processing *function*: $F_I(t) = \sup_{\mathrm{P}_{WX}:\, I(W;X)\leq t} I(W;Y)$



- Also DPI: for any $\mathrm{P}_X, \mathrm{Q}_X$, $D_f(\mathrm{Q}_Y \,\|\, \mathrm{P}_Y) \leq D_f(\mathrm{Q}_X \,\|\, \mathrm{P}_X)$

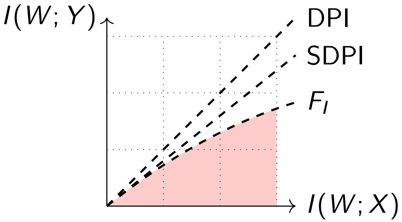
  $\mathrm{P}_Y = \mathrm{P}_X \circ \mathrm{P}_{Y|X}$

Aside on data processing inequalities

Fix $\mathrm{P}_{Y|X}$

- DPI: for any P_{WX}, $I(W;Y) \le I(W;X)$

- Data processing *function*: $F_I(t) = \sup_{\mathrm{P}_{WX}:\, I(W;X) \le t} I(W;Y)$



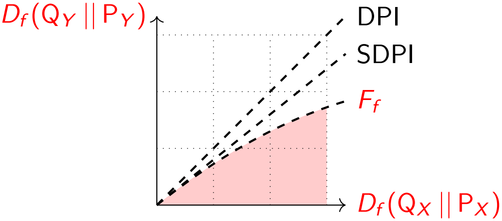
- Also DPI: for any $\mathrm{P}_X, \mathrm{Q}_X$, $D_f(\mathrm{Q}_Y \,||\, \mathrm{P}_Y) \le D_f(\mathrm{Q}_X \,||\, \mathrm{P}_X)$

- Natural analogue: $F_f(t) = \sup_{\mathrm{P}_X, \mathrm{Q}_X:\, D_f(\mathrm{Q}_X \,||\, \mathrm{P}_X) \le t} D_f(\mathrm{Q}_Y \,||\, \mathrm{P}_Y)$

Aside on data processing inequalities

Fix $\mathrm{P}_{Y|X}$

- DPI: for any P_{WX}, $I(W;Y) \le I(W;X)$
- Data processing *function*: $F_I(t) = \sup_{\mathrm{P}_{WX}:\, I(W;X) \le t} I(W;Y)$



- Also DPI: for any $\mathrm{P}_X, \mathrm{Q}_X$, $D_f(\mathrm{Q}_Y \,||\, \mathrm{P}_Y) \le D_f(\mathrm{Q}_X \,||\, \mathrm{P}_X)$
- Natural analogue: $F_f(t) = \sup_{\mathrm{P}_X, \mathrm{Q}_X:\, D_f(\mathrm{Q}_X \,||\, \mathrm{P}_X) \le t} D_f(\mathrm{Q}_Y \,||\, \mathrm{P}_Y)$

Outline

Joint range of input and output divergences

Fix $\mathrm{P}_{Y|X}$ and f

Joint range of input and output divergences

Fix $\mathrm{P}_{Y|X}$ and f

- Define $F_f(t) = \sup_{\mathrm{P}_X, \mathrm{Q}_X : D_f(\mathrm{Q}_X \,||\, \mathrm{P}_X) \leq t} D_f(\mathrm{Q}_Y \,||\, \mathrm{P}_Y)$

Joint range of input and output divergences

Fix $\mathsf{P}_{Y|X}$ and f

- Define $F_f(t) = \sup_{\mathsf{P}_X, \mathsf{Q}_X \colon D_f(\mathsf{Q}_X \,\|\, \mathsf{P}_X) \leq t} D_f(\mathsf{Q}_Y \,\|\, \mathsf{P}_Y)$
- Upper boundary of $\mathcal{D}_f = \bigcup_{\mathsf{P}_X, \mathsf{Q}_X} \left\{ \left(D_f(\mathsf{Q}_X \,\|\, \mathsf{P}_X), D_f(\mathsf{Q}_Y \,\|\, \mathsf{P}_Y) \right) \right\}$

Joint range of input and output divergences

Fix $\mathsf{P}_{Y|X}$ and f

- Define $F_f(t) = \sup_{\mathsf{P}_X, \mathsf{Q}_X \colon D_f(\mathsf{Q}_X \,\|\, \mathsf{P}_X) \le t} D_f(\mathsf{Q}_Y \,\|\, \mathsf{P}_Y)$
- Upper boundary of $\mathcal{D}_f = \bigcup_{\mathsf{P}_X, \mathsf{Q}_X} \left\{ \left(D_f(\mathsf{Q}_X \,\|\, \mathsf{P}_X), D_f(\mathsf{Q}_Y \,\|\, \mathsf{P}_Y) \right) \right\}$
- Conjecture: $\mathcal{D}_f$ is convex ?

Joint range of input and output divergences

Fix $\mathsf{P}_{Y|X}$ and f

- Define $F_f(t) = \sup_{\mathsf{P}_X,\mathsf{Q}_X:\, D_f(\mathsf{Q}_X \,\|\, \mathsf{P}_X) \le t} D_f(\mathsf{Q}_Y \,\|\, \mathsf{P}_Y)$
- Upper boundary of $\mathcal{D}_f = \bigcup_{\mathsf{P}_X,\mathsf{Q}_X} \left\{ \left(D_f(\mathsf{Q}_X \,\|\, \mathsf{P}_X), D_f(\mathsf{Q}_Y \,\|\, \mathsf{P}_Y) \right) \right\}$
- Conjecture: $\mathcal{D}_f$ is convex ($\implies$ F_f is concave) ?

Joint range of input and output divergences

Fix $\mathsf{P}_{Y|X}$ and f

- Define $F_f(t) = \sup_{\mathsf{P}_X, \mathsf{Q}_X : D_f(\mathsf{Q}_X \,\|\, \mathsf{P}_X) \le t} D_f(\mathsf{Q}_Y \,\|\, \mathsf{P}_Y)$
- Upper boundary of $\mathcal{D}_f = \bigcup_{\mathsf{P}_X, \mathsf{Q}_X} \left\{ \left(D_f(\mathsf{Q}_X \,\|\, \mathsf{P}_X), D_f(\mathsf{Q}_Y \,\|\, \mathsf{P}_Y) \right) \right\}$
- Conjecture: $\mathcal{D}_f$ is convex ($\Longrightarrow$ F_f is concave) ?
- Facts:

Joint range of input and output divergences

Fix $\mathrm{P}_{Y|X}$ and f

- Define $F_f(t) = \sup_{\mathrm{P}_X, \mathrm{Q}_X : D_f(\mathrm{Q}_X \,||\, \mathrm{P}_X) \le t} D_f(\mathrm{Q}_Y \,||\, \mathrm{P}_Y)$
- Upper boundary of $\mathcal{D}_f = \bigcup_{\mathrm{P}_X, \mathrm{Q}_X} \left\{ \left(D_f(\mathrm{Q}_X \,||\, \mathrm{P}_X), D_f(\mathrm{Q}_Y \,||\, \mathrm{P}_Y) \right) \right\}$
- Conjecture: $\mathcal{D}_f$ is convex ($\implies$ F_f is concave) ?
- Facts:
  - F_I is NOT necessarily concave (counter-example: $\mathrm{P}_{Y|X} = \mathrm{BEC}^3$)

Joint range of input and output divergences

Fix $\mathsf{P}_{Y|X}$ and f

- Define $F_f(t) = \sup_{\mathsf{P}_X, \mathsf{Q}_X : D_f(\mathsf{Q}_X \,||\, \mathsf{P}_X) \leq t} D_f(\mathsf{Q}_Y \,||\, \mathsf{P}_Y)$
- Upper boundary of $\mathcal{D}_f = \bigcup_{\mathsf{P}_X, \mathsf{Q}_X} \left\{ \left(D_f(\mathsf{Q}_X \,||\, \mathsf{P}_X), D_f(\mathsf{Q}_Y \,||\, \mathsf{P}_Y) \right) \right\}$
- Conjecture: $\mathcal{D}_f$ is convex ($\Longrightarrow$ F_f is concave) ?
- Facts:
  - F_I is NOT necessarily concave (counter-example: $\mathsf{P}_{Y|X} = \mathrm{BEC}^3$)
  - Fix P_X, define $\tilde{F}_I(t, \mathsf{P}_X) = \sup_{\mathsf{P}_{W|X} : I(W;X) \leq t} I(W; Y)$ and

Joint range of input and output divergences

Fix $\mathrm{P}_{Y|X}$ and f

- Define $F_f(t) = \sup_{\mathrm{P}_X, \mathrm{Q}_X : D_f(\mathrm{Q}_X \,||\, \mathrm{P}_X) \le t} D_f(\mathrm{Q}_Y \,||\, \mathrm{P}_Y)$
- Upper boundary of $\mathcal{D}_f = \bigcup_{\mathrm{P}_X, \mathrm{Q}_X} \left\{ \left(D_f(\mathrm{Q}_X \,||\, \mathrm{P}_X), D_f(\mathrm{Q}_Y \,||\, \mathrm{P}_Y) \right) \right\}$
- Conjecture: $\mathcal{D}_f$ is convex ($\implies$ F_f is concave) ?
- Facts:
  - F_I is NOT necessarily concave (counter-example: $\mathrm{P}_{Y|X} = \mathrm{BEC}^3$)
  - Fix P_X, define $\tilde{F}_I(t, \mathrm{P}_X) = \sup_{\mathrm{P}_{W|X} : I(W;X) \le t} I(W;Y)$ and

    $$\tilde{F}_f(t, \mathrm{P}_X) = \sup_{\mathrm{Q}_X : D_f(\mathrm{Q}_X \,||\, \mathrm{P}_X) \le t} D_f(\mathrm{Q}_Y \,||\, \mathrm{P}_Y),$$

Joint range of input and output divergences

Fix $\mathrm{P}_{Y|X}$ and f

- Define $F_f(t) = \sup_{\mathrm{P}_X, \mathrm{Q}_X : D_f(\mathrm{Q}_X \,||\, \mathrm{P}_X) \le t} D_f(\mathrm{Q}_Y \,||\, \mathrm{P}_Y)$
- Upper boundary of $\mathcal{D}_f = \bigcup_{\mathrm{P}_X, \mathrm{Q}_X} \left\{ \left(D_f(\mathrm{Q}_X \,||\, \mathrm{P}_X), D_f(\mathrm{Q}_Y \,||\, \mathrm{P}_Y) \right) \right\}$
- Conjecture: $\mathcal{D}_f$ is convex ($\Longrightarrow$ F_f is concave) ?
- Facts:
  - F_I is NOT necessarily concave (counter-example: $\mathrm{P}_{Y|X} = \mathrm{BEC}^3$)
  - Fix P_X, define $\tilde{F}_I(t, \mathrm{P}_X) = \sup_{\mathrm{P}_{W|X} : I(W;X) \le t} I(W;Y)$ and

    $$\tilde{F}_f(t, \mathrm{P}_X) = \sup_{\mathrm{Q}_X : D_f(\mathrm{Q}_X \,||\, \mathrm{P}_X) \le t} D_f(\mathrm{Q}_Y \,||\, \mathrm{P}_Y),$$

    then $\tilde{F}_I(\cdot, \mathrm{P}_X)$ is concave

Joint range of input and output divergences

Fix $\mathrm{P}_{Y|X}$ and f

- Define $F_f(t) = \sup_{\mathrm{P}_X, \mathrm{Q}_X : D_f(\mathrm{Q}_X \,||\, \mathrm{P}_X) \le t} D_f(\mathrm{Q}_Y \,||\, \mathrm{P}_Y)$
- Upper boundary of $\mathcal{D}_f = \bigcup_{\mathrm{P}_X, \mathrm{Q}_X} \left\{ \left(D_f(\mathrm{Q}_X \,||\, \mathrm{P}_X), D_f(\mathrm{Q}_Y \,||\, \mathrm{P}_Y) \right) \right\}$
- Conjecture: $\mathcal{D}_f$ is convex ($\implies$ F_f is concave) ?
- Facts:
  - F_I is NOT necessarily concave (counter-example: $\mathrm{P}_{Y|X} = \mathrm{BEC}^3$)
  - Fix P_X, define $\tilde{F}_I(t, \mathrm{P}_X) = \sup_{\mathrm{P}_{W|X} : I(W;X) \le t} I(W;Y)$ and

$$\tilde{F}_f(t, \mathrm{P}_X) = \sup_{\mathrm{Q}_X : D_f(\mathrm{Q}_X \,||\, \mathrm{P}_X) \le t} D_f(\mathrm{Q}_Y \,||\, \mathrm{P}_Y),$$

    then $\tilde{F}_I(\cdot, \mathrm{P}_X)$ is concave; $\implies$ $\tilde{F}_f(\cdot, \mathrm{P}_X)$ is concave

Joint range of input and output divergences

Fix $\mathrm{P}_{Y|X}$ and f

- Define $F_f(t) = \sup_{\mathrm{P}_X, \mathrm{Q}_X : D_f(\mathrm{Q}_X \| \mathrm{P}_X) \le t} D_f(\mathrm{Q}_Y \| \mathrm{P}_Y)$
- Upper boundary of $\mathcal{D}_f = \bigcup_{\mathrm{P}_X, \mathrm{Q}_X} \left\{ \left(D_f(\mathrm{Q}_X \| \mathrm{P}_X), D_f(\mathrm{Q}_Y \| \mathrm{P}_Y) \right) \right\}$
- Conjecture: $\mathcal{D}_f$ is convex ($\implies$ F_f is concave) ?
- Facts:
  - F_I is NOT necessarily concave (counter-example: $\mathrm{P}_{Y|X} = \mathrm{BEC}^3$)
  - Fix P_X, define $\tilde{F}_I(t, \mathrm{P}_X) = \sup_{\mathrm{P}_{W|X} : I(W;X) \le t} I(W;Y)$ and
    $$\tilde{F}_f(t, \mathrm{P}_X) = \sup_{\mathrm{Q}_X : D_f(\mathrm{Q}_X \| \mathrm{P}_X) \le t} D_f(\mathrm{Q}_Y \| \mathrm{P}_Y),$$
    then $\tilde{F}_I(\cdot, \mathrm{P}_X)$ is concave; $\implies$ $\tilde{F}_f(\cdot, \mathrm{P}_X)$ is concave
  - For any f, g, $\bigcup_{\mathrm{P}_X, \mathrm{Q}_X} \left\{ \left(D_f(\mathrm{Q}_X \| \mathrm{P}_X), D_g(\mathrm{Q}_X \| \mathrm{P}_X) \right) \right\}$ is convex

Outline

1. Prompt compression for black-box language models
2. Input-entropy-constrained capacity
3. Joint range of divergences
4. Closing remarks

In closing. . .

In closing...

- Three problems ($1\times$ compression, $2\times$ contraction):

In closing. . .

- Three problems ($1\times$ compression, $2\times$ contraction):
  - Prompt compression for LLMs

In closing...

- Three problems ($1\times$ compression, $2\times$ contraction):
  - Prompt compression for LLMs
  - Entropy-constrained capacity

In closing...

- Three problems ($1\times$ compression, $2\times$ contraction):
  - Prompt compression for LLMs
  - Entropy-constrained capacity
  - Joint range of divergences

In closing...

- Three problems ($1\times$ compression, $2\times$ contraction):
  - Prompt compression for LLMs
  - Entropy-constrained capacity
  - Joint range of divergences
- Two more:

In closing. . .

- Three problems ($1\times$ compression, $2\times$ contraction):
  - Prompt compression for LLMs
  - Entropy-constrained capacity
  - Joint range of divergences
- Two more:
  - Guesswork

In closing...

- Three problems ($1\times$ compression, $2\times$ contraction):
    - Prompt compression for LLMs
    - Entropy-constrained capacity
    - Joint range of divergences
- Two more:
    - Guesswork
    - Distributed hypothesis testing

In closing. . .

- Three problems ($1\times$ compression, $2\times$ contraction):
  - Prompt compression for LLMs
  - Entropy-constrained capacity
  - Joint range of divergences
- Two more (method of types + optimization):
  - Guesswork
  - Distributed hypothesis testing

In closing...

- Three problems ($1\times$ compression, $2\times$ contraction):
  - Prompt compression for LLMs
  - Entropy-constrained capacity
  - Joint range of divergences
- Two more (method of types $+$ optimization):
  - Guesswork
  - Distributed hypothesis testing $\longrightarrow$ compression $+$ contraction

In closing...

- Three problems ($1\times$ compression, $2\times$ contraction):
  - Prompt compression for LLMs
  - Entropy-constrained capacity
  - Joint range of divergences
- Two more (method of types + optimization):
  - Guesswork
  - Distributed hypothesis testing $\longrightarrow$ compression + contraction
- All thoughts welcome

In closing...

- Three problems ($1\times$ compression, $2\times$ contraction):
  - Prompt compression for LLMs
  - Entropy-constrained capacity
  - Joint range of divergences
- Two more (method of types + optimization):
  - Guesswork
  - Distributed hypothesis testing $\longrightarrow$ compression + contraction
- All thoughts welcome

Thank you!