Variations on a Theme by Liu, Cuff, and Verdú

The Power of Posterior Sampling

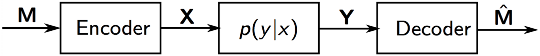
Alankrita Bhatt, Jiun-Ting Huang, Young-Han Kim, J. Jon Ryu, and Pinar Sen

Department of Electrical and Computer Engineering
University of California, San Diego

IEEE Information Theory Workshop
Guangzhou, China
November 27, 2018

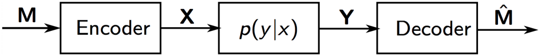
Aria

- **Problem**: Detect a signal X from an observation Y to minimize $P_e = \mathrm{P}\{X \neq \hat{x}(Y)\}$
- **Example**: In channel coding, find $\hat{\mathbf{m}}(\mathbf{Y})$ that minimizes $\mathrm{P}\{\hat{\mathbf{m}}(\mathbf{Y}) \neq \mathbf{M}\}$



Aria

- **Problem**: Detect a signal X from an observation Y to minimize $P_e = \mathrm{P}\{X \neq \hat{x}(Y)\}$
- **Example**: In channel coding, find $\hat{\mathbf{m}}(\mathbf{Y})$ that minimizes $\mathrm{P}\{\hat{\mathbf{m}}(\mathbf{Y}) \neq \mathbf{M}\}$

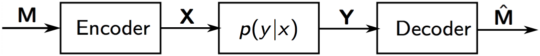


- Optimal detector: Maximum a-posteriori probability (MAP)

$$\hat{x}^*(y) = \arg\max_{x \in \mathcal{X}} p(x|y)$$

Aria

- **Problem**: Detect a signal X from an observation Y to minimize $P_e = \mathrm{P}\{X \neq \hat{x}(Y)\}$
- **Example**: In channel coding, find $\hat{\mathbf{m}}(\mathbf{Y})$ that minimizes $\mathrm{P}\{\hat{\mathbf{m}}(\mathbf{Y}) \neq \mathbf{M}\}$



- Optimal detector: Maximum a-posteriori probability (MAP)

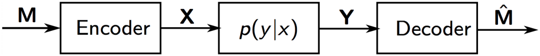
$$\hat{x}^*(y) = \arg\max_{x \in \mathcal{X}} p(x|y)$$

- Randomized likelihood (RL) detector (Yassaee et al. 2013):

$$\hat{X}(y) \sim p(x|y)$$

Aria

- **Problem**: Detect a signal X from an observation Y to minimize $P_e = \mathsf{P}\{X \neq \hat{x}(Y)\}$
- **Example**: In channel coding, find $\hat{\mathbf{m}}(\mathbf{Y})$ that minimizes $\mathsf{P}\{\hat{\mathbf{m}}(\mathbf{Y}) \neq \mathbf{M}\}$



- Optimal detector: Maximum a-posteriori probability (MAP)

$$\hat{x}^*(y) = \arg\max_{x \in \mathcal{X}} p(x|y)$$

- Randomized likelihood (RL) detector (Yassaee et al. 2013):

$$\hat{X}(y) \sim p(x|y)$$

Liu–Cuff–Verdú lemma (2017)

If $\hat{X}$ is a conditionally i.i.d. copy of X given Y, then

$$\mathsf{P}\{X \neq \hat{X}\} \leq 2P_e^* = 2\,\mathsf{P}\{X \neq \hat{x}^*(Y)\}$$

Proof of the LCV lemma

A more general inequality

For any metric $d(x, x')$ and $X \overset{d}{=} X'$,

$$\mathsf{E}[d(X, X')] \le 2 \inf_x \mathsf{E}[d(X, x)]$$

- *Proof:* Since $d(x, x') \le d(x, x'') + d(x', x'')$ and $X \overset{d}{=} X'$

$$\mathsf{E}[d(X, X')] \le \mathsf{E}[d(X, x'')] + \mathsf{E}[d(X', x'')] = 2\,\mathsf{E}[d(X, x'')], \quad \forall x''$$

  Now take the infimum over x'' on both sides

Proof of the LCV lemma

A more general inequality

For any metric $d(x, x')$ and $X \stackrel{d}{=} X'$,

$$\mathsf{E}[d(X, X')] \leq 2 \inf_x \mathsf{E}[d(X, x)]$$

- *Proof:* Since $d(x, x') \leq d(x, x'') + d(x', x'')$ and $X \stackrel{d}{=} X'$

  $$\mathsf{E}[d(X, X')] \leq \mathsf{E}[d(X, x'')] + \mathsf{E}[d(X', x'')] = 2\,\mathsf{E}[d(X, x'')], \quad \forall x''$$

  Now take the infimum over x'' on both sides

- *Proof of the LCV lemma:* For $d(x, \hat{x}) = \mathbb{1}_{\{x \neq \hat{x}\}}$ and $X \stackrel{d}{=} \hat{X} | \{Y = y\}$

  $$\mathsf{P}\{X \neq \hat{X} | Y = y\} \leq 2 \inf_x \mathsf{P}\{X \neq x | Y = y\} = 2\,\mathsf{P}\{X \neq \hat{x}^*(y) | Y = y\}$$

  Now take expectation w.r.t Y on both sides

Historical remarks

- The current form due to Jingbo Liu (ca Spring 2015)
- A similar observation also made by Kudekar et al. (2016)

Historical remarks

- The current form due to Jingbo Liu (ca Spring 2015)
- A similar observation also made by Kudekar et al. (2016)
- Channel coding: The RL decoder $\hat{\mathbf{M}}_{\mathrm{RL}} \sim p(\mathbf{m}|\mathbf{Y})$ achieves

$$\mathsf{P}\{\mathbf{M} \neq \hat{\mathbf{M}}\} = P_e \leq 2P_e^* = 2\,\mathsf{P}\{\mathbf{M} \neq \hat{\mathbf{m}}^*(\mathbf{Y})\}$$

  Similarly for the bit error rate

$$\frac{1}{n}\sum_{i=1}^{n} \mathsf{P}\{M_i \neq \hat{M}_i\} = P_b \leq 2P_b^* = 2\left(\frac{1}{n}\sum_{i=1}^{n} \mathsf{P}\{M_i \neq \hat{m}_i^*(\mathbf{Y})\}\right)$$

- Monte Carlo decoding: Efficiently sampling from the posterior $p(\mathbf{m}|\mathbf{Y})$ (Neal 2001, Mezard–Montanari 2009, Bhatt et al. 2018)

Historical remarks

- The current form due to Jingbo Liu (ca Spring 2015)
- A similar observation also made by Kudekar et al. (2016)
- Channel coding: The RL decoder $\hat{\mathbf{M}}_{\mathrm{RL}} \sim p(\mathbf{m}|\mathbf{Y})$ achieves

  $$\mathsf{P}\{\mathbf{M} \neq \hat{\mathbf{M}}\} = P_e \leq 2P_e^* = 2\,\mathsf{P}\{\mathbf{M} \neq \hat{\mathbf{m}}^*(\mathbf{Y})\}$$

  Similarly for the bit error rate

  $$\frac{1}{n}\sum_{i=1}^{n} \mathsf{P}\{M_i \neq \hat{M}_i\} = P_b \leq 2P_b^* = 2\left(\frac{1}{n}\sum_{i=1}^{n} \mathsf{P}\{M_i \neq \hat{m}_i^*(\mathbf{Y})\}\right)$$

- Monte Carlo decoding: Efficiently sampling from the posterior $p(\mathbf{m}|\mathbf{Y})$ (Neal 2001, Mezard–Montanari 2009, Bhatt et al. 2018)
- Original $2\times$ bound: Cover–Hart theorem (1967) for nearest-neighbor classification

Variation 1

- This talk presents 9 variations of the LCV lemma

Variation 1

- This talk presents 9 variations of the LCV lemma
- As a warm-up, the first variation strengthens the LCV lemma when $|\mathcal{X}| < \infty$

Tighter LCV lemma

If $\mathcal{X}$ is finite,

$$\mathsf{P}\{X \neq X'\} \leq 2P_e^*\left(1 - \frac{|\mathcal{X}|}{2(|\mathcal{X}|-1)}P_e^*\right).$$

- *Proof:* Cauchy–Schwartz and Jensen (which traces back to Cover and Hart)

Variation 2

- Let $\mathcal{S}_{y,N} = \{X'_1, X'_2, \cdots, X'_N\}$ be a set of conditionally i.i.d. copies of X given Y

Variation 2

- Let $\mathcal{S}_{y,N} = \{X'_1, X'_2, \cdots, X'_N\}$ be a set of conditionally i.i.d. copies of X given Y
- Let
$$q_1(y) = \max\{p(x|y) : x \in \mathcal{X}\}$$
and
$$\epsilon(\delta) = \mathsf{P}\{q_1(Y) \le \delta\}$$

Empirical MAP

Let
$$\hat{X}_N = \arg\max_{\mathcal{S}_{y,N}} p(x|y)$$
Then
$$\mathsf{P}\{X \neq \hat{X}_N\} \le P_e^* + e^{-\delta N} + \epsilon(\delta)$$

- *Proof:* $\mathsf{P}\{\hat{X}_N \notin \mathcal{X}^*(y) | Y = y\} \le e^{-q_1(y)N}$, where $\mathcal{X}^*(y) = \{x \in \mathcal{X} : p(x|y) = q_1(y)\}$

Variation 3

- Let

$$c_3 = \max_{p \in [0,1]} (1 + 3p - 5p^2 + 2p^3) \approx 1.528$$
$$c_5 = \max_{p \in [0,1]} (1 + 10p^2 - 25p^3 + 21p^4 - 6p^5) \approx 1.501$$

Empirical mode #1

Let

$$\hat{X}_N = \mathrm{mode}(X'_1, X'_2, \cdots, X'_N)$$

Then

$$\mathsf{P}\{X \neq \hat{X}_3\} \leq c_3 P_e^*$$
$$\mathsf{P}\{X \neq \hat{X}_5\} \leq c_5 P_e^*$$

- *Proof:* A careful analysis of the majority vote

Variation 4

- Let

$$q_1(y) = \max\{p(x|y) \colon x \in \mathcal{X}\}$$
$$q_2(y) = \max\{p(x|y) \colon x \in \mathcal{X} \backslash \mathcal{X}^*(y)\}$$

  where $\mathcal{X}^*(y) = \{x \in \mathcal{X} \colon p(x|y) = q_1(y)\}$

- Let

$$\Delta(y) = q_1(y) - q_2(y)$$
$$\epsilon(\delta) = \mathsf{P}\{\Delta(Y) \leq \delta\}$$

Variation 4

- Let

$$q_1(y) = \max\{p(x|y) : x \in \mathcal{X}\}$$
$$q_2(y) = \max\{p(x|y) : x \in \mathcal{X} \setminus \mathcal{X}^*(y)\}$$

where $\mathcal{X}^*(y) = \{x \in \mathcal{X} : p(x|y) = q_1(y)\}$

- Let

$$\Delta(y) = q_1(y) - q_2(y)$$
$$\epsilon(\delta) = \mathsf{P}\{\Delta(Y) \le \delta\}$$

Empirical mode #2

For any $\delta > 0$

$$\mathsf{P}\{X \ne \hat{X}_N\} \le P_e^* + \min\{(|\mathcal{X}| - 1)(e^{-\frac{\delta^2 N}{2}} + \epsilon(\delta)), 8(N+1)(e^{-\frac{\delta^2 N}{128}} + \epsilon(\delta))\}$$

- *Proof:* Hoeffding and Vapnik–Chervonenkis

Interlude

- So far we discussed detection

Interlude

- So far we discussed detection
- **Problem:** Estimate a signal X from an observation Y to minimize $\mathsf{E}[d(X, \hat{x}(Y))]$
- **Examples:** Minimum MAE and MSE estimation

Interlude

- So far we discussed detection
- **Problem:** Estimate a signal X from an observation Y to minimize $\mathsf{E}[d(X, \hat{x}(Y))]$
- **Examples:** Minimum MAE and MSE estimation
- Optimal estimator: Bayes

$$\hat{x}^*(y) = \inf_{\hat{x} \in \mathcal{X}} \mathsf{E}[d(X, \hat{x}(y)) | Y = y]$$

- Randomized likelihood estimator: $\hat{X} \sim f(x|y)$

Variation 5

- Absolute loss: $d(x, \hat{x}) = |x - \hat{x}|$
- Optimal estimator: Conditional median

$$\hat{x}^*(y) = \inf_x \{x \colon F(x|y) \geq 1/2\}$$

Variation 5

- Absolute loss: $d(x, \hat{x}) = |x - \hat{x}|$
- Optimal estimator: Conditional median

$$\hat{x}^*(y) = \inf_x \{x : F(x|y) \geq 1/2\}$$

Mean absolute error (MAE) estimation

If $\hat{X}$ is a conditionally i.i.d. copy of X given Y, then

$$\mathsf{E}[|X - \hat{X}|] \leq 2\,\mathsf{E}[|X - \hat{x}^*(Y)|]$$

- *Proof:* Recall $\mathsf{E}[d(X, X')] \leq 2 \inf_x \mathsf{E}[d(X, x)]$

Variation 6

- Let $X_1', X_2', \ldots, X_N'$ be conditionally i.i.d. copies of X given Y
- Let

$$\hat{F}_N(x|y) := \frac{1}{N}\sum_{i=1}^{N} 1_{\{X_i(y)\leq x\}}$$

Variation 6

- Let $X_1', X_2', \ldots, X_N'$ be conditionally i.i.d. copies of X given Y
- Let

$$\hat{F}_N(x|y) := \frac{1}{N} \sum_{i=1}^{N} 1_{\{X_i(y) \le x\}}$$

Empirical median

Let

$$\hat{X}_N = \inf_x \{x \colon \hat{F}_N(x|y) \ge 1/2\}$$

If $|X| \le B < \infty$ a.s. and $|F(\hat{x}^*(y) + \alpha|y) - 1/2| \ge L|\alpha|$ for $\alpha \in (-r, r)$, then

$$\mathsf{E}[|X - \hat{X}_N|] \le \mathsf{E}[|X - \hat{x}^*(Y)|] + 4Be^{-2L^2(\epsilon \wedge r)^2 N} + \epsilon$$

- *Proof*: Dvoretzky–Kiefer–Wolfowitz

Variation 7

- Quadratic loss $d(x, \hat{x}) = (x - \hat{x})^2$
- Optimal estimator: Conditional mean (expectation)

$$\hat{x}^*(y) = \mathsf{E}[X | Y = y].$$

Variation 7

- Quadratic loss $d(x, \hat{x}) = (x - \hat{x})^2$
- Optimal estimator: Conditional mean (expectation)

$$\hat{x}^*(y) = \mathsf{E}[X | Y = y].$$

Mean squared error (MSE) estimation

If $\hat{X}$ is a conditionally i.i.d. copy of X given Y, then

$$\mathsf{E}[(X - \hat{X})^2] = 2\,\mathsf{E}[(X - \hat{x}^*(Y))^2]$$

- *Proof:* Although $d(x, x') = (x - x')^2$ is not a metric, if $X \stackrel{d}{=} X'$, then

$$\mathsf{E}[(X - X')^2] = \mathsf{E}[(X - \mu - X' + \mu)^2] = \mathsf{E}[(X - \mu)^2] + \mathsf{E}[(X' - \mu)^2]$$

Variations 8 & 9

- Let $X_1', X_2', \ldots, X_N'$ be conditionally i.i.d. copies of X given Y

Empirical mean #1

Let

$$\hat{X}_N = \frac{1}{N}(X_1' + \cdots + X_N')$$

Then

$$\mathsf{E}[(X - \hat{X}_N)^2] = \left(1 + \frac{1}{N}\right)\mathsf{E}[(X - \hat{x}^*(Y))^2]$$

- *Proof:* Algebra

Variations 8 & 9

- Let $X_1', X_2', \ldots, X_N'$ be conditionally i.i.d. copies of X given Y

Empirical mean #1

Let

$$\hat{X}_N = \frac{1}{N}(X_1' + \cdots + X_N')$$

Then

$$\mathsf{E}[(X - \hat{X}_N)^2] = \left(1 + \frac{1}{N}\right)\mathsf{E}[(X - \hat{x}^*(Y))^2]$$

- *Proof:* Algebra

Empirical mean #2

If $|X| \leq B < \infty$ a.s., then

$$\mathsf{P}\{|X - \hat{X}_N| \geq \epsilon\} \leq \mathsf{P}\{|X - \hat{x}^*(Y)| \geq (1-\delta)\epsilon\} + 2e^{-\delta^2\epsilon^2 N/4B^2}$$

- *Proof:* Hoeffding

Coda

- The factor-of-two bound from posterior sampling is ubiquitous

Coda

- The factor-of-two bound from posterior sampling is ubiquitous
- Taking multiple samples often converges essentially exponentially to Bayes

Coda

- The factor-of-two bound from posterior sampling is ubiquitous

- Taking multiple samples often converges essentially exponentially to Bayes

- **Open problem #1:** Tighter and more general bounds?

- **Open problem #2:** Applications, applications, and applications?

References

Bhatt, A., Huang, J.-T., Kim, Y.-H., Ryu, J. J., and Sen, P. (2018). Monte Carlo methods for randomized likelihood decoding. In *Proc. 56th Ann. Allerton Conf. Comm. Control Comput.*, Monticello, IL.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory*, 13(1), 21–27.

Kudekar, S., Kumar, S., Mondelli, M., Pfister, H., and Urbanke, R. (2016). Comparing the bit-MAP and block-MAP decoding thresholds of Reed-Muller codes on BMS channels. In *Proc. IEEE Int. Symp. Inf. Theory*, Barcelona, Spain, pp. 1755–1759.

Liu, J., Cuff, P., and Verdú, S. (2017). On α-decodability and α-likelihood decoder. In *Proc. 55th Ann. Allerton Conf. Comm. Control Comput.*, Monticello, IL.

Mezard, M. and Montanari, A. (2009). *Information, Physics, and Computation*. Oxford University Press.

Neal, R. M. (2001). Monte Carlo decoding of LDPC codes. Tech. rep., Dept. of Computer Science, University of Toronto. Presented at *ICTP Works. Statis. Phys. Capacity-Approaching Codes*.

Yassaee, M. H., Aref, M. R., and Gohari, A. (2013). A technique for deriving one-shot achievability results in network information theory. In *Proc. IEEE Int. Symp. Inf. Theory*, Istanbul, Turkey, pp. 1151–1155.