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

***Abstract*—The Liu–Cuff–Verdú lemma states that in estimating a source X from an observation Y, making a random guess X' from the posterior $p(x|y)$ can go wrong at most twice as often as the optimal answer. Several variations of this fundamental, yet rather arcane, result are explored for detection, decoding, and estimation problems.**

I. Introduction

Optimal detection is one of the most fundamental problems in statistical signal processing. Suppose that a discrete signal $X \in \mathcal{X}$ is to be estimated from a noisy observation Y. Which estimate $\hat{X} = \hat{x}(Y)$ minimizes the probability of error $P_e = \mathsf{P}\{X \neq \hat{X}\}$? The same problem arises for optimal decision making or optimal decoding of error correcting codes. For example, a student is taking a multiple-choice test and has to make a decision on which answer to select from A, B, C, and D. When the conditional probability (or the posterior in the Bayesian parlance) $p(x|y)$ of the signal (the true answer) given the observation (the student's understanding of the question) is known, then the answer to the optimal detection problem is straightforward. One should simply choose $\hat{x}(y)$ that maximizes the posterior $p(x|y)$ for each realization y of the observation Y. This optimal detector

$$\hat{x}^*(y) = \arg\max_{x \in \mathcal{X}} p(x|y) \tag{1}$$

is often referred to as the maximum a posterior probability (MAP) detector, with ties in the maximum (if any) broken in an arbitrary way. No other detector can achieve the probability of error smaller than the minimum

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

achieved by the MAP detector. If the answer to the question is A with probability 10%, B with probability 20%, C with probability 30%, and D with probability 40%, then the student should choose D to minimize the probability of error as 60%.

Now what happens if one succumbs to a sudden whim and instead makes a random choice according to the posterior probability, say, $(.1, .2, .3, .4)$? The following observation by Jingbo Liu, which was related to one of the authors by his then Ph.D. advisor Paul Cuff during the 2015 IEEE ISIT in Hong Kong, asserts that a randomly generated answer from the posterior is not too far off from the optimal MAP decision.

Liu–Cuff–Verdú lemma (published in [1])**.** *Let X' be a conditionally independent and identically distributed (i.i.d.) copy of X given Y, i.e., $X'|\{Y = y\} \sim p(x|y)$. Then*

$$P_e^* \leq \mathsf{P}\{X \neq X'\} \leq 2P_e^*.$$

We prove the Liu–Cuff–Verdú (LCV) lemma through the following abstraction.

Lemma 1. *Let $d\colon \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be a metric defined over an alphabet $\mathcal{X}$. Let X and X' be i.i.d. random variables on $\mathcal{X}$. Then*

$$\mathsf{E}[d(X, X')] \leq 2 \inf_{x \in \mathcal{X}} \mathsf{E}[d(X, x)]. \tag{2}$$

In particular,

$$\mathsf{P}\{X \neq X'\} \leq 2 \inf_{x \in \mathcal{X}} \mathsf{P}\{X \neq x\}. \tag{3}$$

Proof: Since

$$d(x, x') \leq d(x, x'') + d(x', x'')$$

for any x, x', x'', and X and X' are identically distributed,

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

for any $x \in \mathcal{X}$. Taking infimum over $x \in \mathcal{X}$ on both sides establishes (2), which can be specialized with the Hamming distortion $d(x, x') = \mathbb{1}\{x \neq x'\}$ to establish (3). ∎

Proof of the LCV lemma: The lower bound follows trivially. By the conditional version of Lemma 1, since X and X' are conditionally i.i.d.,

$$\begin{aligned}\mathsf{P}\{X \neq X'|Y = y\} &\leq 2 \inf_{x \in \mathcal{X}} \mathsf{P}\{X \neq x|Y = y\} \\ &= 2\,\mathsf{P}\{X \neq \hat{x}^*(y)|Y = y\}.\end{aligned}$$

Taking expectation with respect to (w.r.t.) Y on both sides establishes the upper bound. ∎

The LCV lemma takes its root in the classical result by Cover and Hart [2], which shows that the error probability of the nearest-neighbor classifier is at most twice the Bayes optimal error probability in the sample limit. In the context of optimal decoding in communication, the paper [1] by Liu, Cuff, and Verdú seems to be the first in the literature to point out the factor-of-two bound on the error probability of the randomized likelihood decoder. The Cover–Hart analysis of multilabel nearest-neighbor classification can be applied to strengthen the LCV lemma as follows.

Variation 1. *If $|\mathcal{X}|$ is finite, then*

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

Beyond the power of random guesses in a test, Liu–Cuff–Verdú lemma has the following implications on standard signal processing and communication problems.

978-1-5386-3599-5/18/$31.00©2018 IEEE

Example 1 (Array signal processing)**.** Consider the optimal detection problem in a multiple-input multiple-output (MIMO) system with input $\mathbf{X}$ and the output

$$\mathbf{Y} = \mathsf{H}\mathbf{X} + \mathbf{Z},$$

where $\mathbf{Z} \sim \mathrm{N}(0, I_n)$ and $\mathsf{H} \in \mathbb{R}^{n\times n}$. Typically $\mathbf{x}$ takes values from a subset of a lattice and optimal detection reduces to integer least-squares (ILS) problem [3]. Sphere decoding (see, for example, [4], [5]) is an efficient algorithm for the ILS problem, but when n is large, its complexity is still prohibitively high. To overcome this difficulty, several low-complexity alternative methods [6], [7] has been proposed based on Markov chain Monte Carlo (MCMC) sampling. The LCV lemma provides theoretical performance guarantee for these algorithms (even with a single sample). Even when the sample is drawn from an approximate distribution $q(x|y)$ instead of the true posterior $p(x|y)$, we have

$$\mathsf{P}\{X \neq X'\} \le 2P_e^* + 2\,\mathsf{E}[d_{\mathrm{TV}}(p(x|Y), q(x|Y))].$$

Example 2 (Channel coding)**.** The LCV lemma is also applicable to the point-to-point communication problem in which a k-bit message $\mathbf{M} = (M_1, \ldots, M_k)$ is encoded into a codeword $x(\mathbf{M})$ and transmitted over a communication channel $p(y|x)$. If $\mathbf{M}$ is drawn uniformly at random, the optimal MAP decoder is equivalent to the maximum likelihood (ML) decoder

$$\hat{\mathbf{m}}^*(y) = \arg\max_{\mathbf{m}} p(y|x(\mathbf{m})).$$

In their award-winning paper, Yassaee, Gohari, and Aref [8] proposed the (randomized) likelihood decoder that generates a sample $\mathbf{M}'$ from the posterior $p(\mathbf{m}|y)$ as a proof device for nonasymptotic coding theorems in information theory. A dual technique of likelihood encoding was also used for lossy source coding [9]. The LCV lemma shows that there is more to the likelihood decoder. On the theoretical side, as pointed out in [1], The LCV lemma provides elementary and much stronger proofs for the existing analyses [10], [11] of the performance of the likelihood decoder (such as its achievable error exponent). On the practical side, the LCV lemma provides theoretical justifications for heuristic approaches to decoding of LDPC codes explored by Neal [12], and Mezard and Montanari [13] based on MCMC methods (see also [14]). Lemma 1 can be also applied immediately to the bit-error rate (BER)

$$d(\mathbf{m}, \mathbf{m}') = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}_{\{\mathbf{m}_i \neq \mathbf{m}'_i\}}.$$

Hence, the same likelihood decoder $\mathbf{M}' \sim p(\mathbf{m}|y)$ also enjoys the factor-of-two performance guarantee from the minimum BER of the optimal bit ML decoder. Note that if $\mathbf{M}'_\alpha \sim \frac{p^\alpha(\mathbf{m}|y)}{\sum_{\overline{\mathbf{m}}} p^\alpha(\overline{\mathbf{m}}|y)}$, then $\mathsf{P}\{\mathbf{M}'_\alpha \neq \mathbf{M}\}$ monotonically decreases for $\alpha \ge 1$ and approaches P_e^* as $\alpha \to \infty$ [1, Theorem 3], and hence that the exponentiated likelihood decoder $\mathbf{M}'_\alpha$ also has the factor-of-two performance guarantee for $\alpha \ge 1$ (cf. [15]).

Motivated by these examples, this paper explores other implications of posterior sampling in statistical signal processing. Our development traverses two different directions. As with the performance guarantee of posterior sampling for block and bit error rates, we show that a single sample from the posterior has a similar factor-of-two performance guarantee for optimal estimation under mean absolute and square error criteria in Section III. Following the empirical mode decoder in [13], we also investigate the case in which multiple samples are drawn from the posterior. For each of optimal detection/decoding, minimum mean absolute error (MMAE) estimation, and minimum mean square error (MMSE) estimation, we show that the factor of two can be improved essentially to the factor of one in the sample limit. Due to the space limitation, proofs of some technical results will be omitted.

II. Detection Based on Multiple Samples

A. MAP Among the Samples

Let

$$\mathcal{S} = \{X'_1, X'_2, \ldots, X'_N\}$$

be a set of random samples from the posterior $p(x|y)$, that is, $X'_1, \ldots, X'_N$ are conditionally i.i.d. copies of X given Y. Let

$$\hat{X}_N = \arg\max_{x\in\mathcal{S}} p(x|y) \tag{4}$$

be the maximum a posteriori probability sample among the N random samples in $\mathcal{S}$. As in the MAP detector in (1), ties are broken in an arbitrary way as before. When $N = 1$, this MAP-among-the-samples detector reduces to X' in Liu–Cuff–Verdú lemma. When $|\mathcal{X}|$ is finite, then $\mathcal{S}$ would ultimately converge to $\mathcal{X}$ as $N \to \infty$, and consequently $\mathsf{P}\{X \neq \hat{X}_N\}$ should converge to the minimum error probability P_e^*.

The following result shows that the convergence is essentially exponential, even for a countably infinite alphabet $\mathcal{X}$.

Variation 2. *Let* $q_1(y) \coloneqq \max\{p(x|y)\colon x \in \mathcal{X}\}$ *and*

$$\epsilon(\delta) = \mathsf{P}\{q_1(Y) \le \delta\} \tag{5}$$

Then, for any $\delta > 0$*,*

$$\mathsf{P}\{X \neq \hat{X}_N\} \le P_e^* + e^{-N\delta} + \epsilon(\delta)$$

In particular, if $\operatorname{ess\,inf}_{y\in\mathcal{Y}} q_1(y) > 0$*,* $\mathsf{P}\{X \neq \hat{X}_N\}$ *converges to* P_e^* *exponentially in* N*.*

Our proof uses the following.

Lemma 2. *Let*

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

Then, for every $y \in \mathcal{Y}$

$$\mathsf{P}\{\hat{X}_N \notin \mathcal{X}^*(y) | Y = y\} \le e^{-q_1(y)N} \tag{6}$$

Proof: We have $\hat{X}_N \notin \mathcal{X}^*(y)$ if and only if $X'_i \notin \mathcal{X}^*(y)$ for all $i \in [N]$. Thus

$$\begin{aligned}\mathsf{P}\{\hat{X}_N \notin \mathcal{X}^*(y) | Y = y\} &= (1 - |\mathcal{X}^*(y)|q_1(y))^N \\ &\le e^{-|\mathcal{X}^*(y)|q_1(y)N},\end{aligned}$$

where the inequality follows since $\ln z \le z - 1$. Finally, noting that $|\mathcal{X}^*(y)| \ge 1$ establishes (6). ∎

Proof of Variation 2: Assume without loss of generality that the MAP detector in (1) breaks ties uniformly at random and denote this (random) optimal detector as $\hat{X}^*$. For each symbol $x \in \mathcal{X}$, define

$$E_0(x) \coloneqq \{y\colon x \notin \mathcal{X}^*(y)\},$$
$$E_k(x) \coloneqq \{y\colon x \in \mathcal{X}^*(y), |\mathcal{X}^*(y)| = k\}, \quad k \in \mathbb{N}.$$

Then, for any x and y,

$$\begin{aligned}
&\mathsf{P}\{X \neq \hat{X}^* | Y = y, X = x\} \\
&= \mathsf{P}\{\hat{X}^* \neq x | Y = y\} \\
&= \mathbb{1}_{\{y \in E_0(x)\}} + \sum_{k=1}^{|\mathcal{X}|} \mathsf{P}\{\hat{X}^* \neq x | Y = y\} \mathbb{1}_{\{y \in E_k(x)\}} \\
&= \mathbb{1}_{\{y \in E_0(x)\}} + \sum_{k=1}^{|\mathcal{X}|} \frac{k-1}{k} \mathbb{1}_{\{y \in E_k(x)\}}. \qquad (7)
\end{aligned}$$

Now for any detector $\hat{X}_N$ (not necessarily the one in (4)),

$$\begin{aligned}
&\mathsf{P}\{X \neq \hat{X}_N | Y = y, X = x\} \\
&= \sum_{k=0}^{|\mathcal{X}|} \mathsf{P}\{\hat{X}_N \neq x | Y = y\} \mathbb{1}_{\{y \in E_k(x)\}} \\
&\leq \mathbb{1}_{\{y \in E_0(x)\}} + \sum_{k=1}^{|\mathcal{X}|} \frac{k-1}{k} \mathbb{1}_{\{y \in E_k(x)\}} \\
&\quad + \sum_{k=1}^{|\mathcal{X}|} \frac{1}{k} \mathsf{P}\{\hat{X}_N \notin \mathcal{X}^*(y) | Y = y\} \mathbb{1}_{\{y \in E_k(x)\}} \qquad (8) \\
&\leq \mathsf{P}\{X \neq \hat{X}^* | Y = y\} + \mathsf{P}\{\hat{X}_N \notin \mathcal{X}^*(y) | Y = y\}. \qquad (9)
\end{aligned}$$

To justify (8), suppose that $y \in E_k(x)$ for $k \geq 1$. Then,

$$\mathsf{P}\{\hat{X}_N \in \mathcal{X}^*(y) | Y = y\} = k\, \mathsf{P}\{\hat{X}_N = x | Y = y\},$$

which implies

$$\mathsf{P}\{\hat{X}_N \neq x | Y = y\} = \frac{k-1}{k} + \frac{1}{k} \mathsf{P}\{\hat{X}_N^1 \notin \mathcal{X}^*(y) | Y = y\}.$$

Also, (9) follows by recalling (7) and replacing $1/k$ with 1 in (8). Taking expectation on both sides of (9), and using Lemma 2 and (5), we have the desired result. ∎

B. Empirical Mode

The MAP-among-the-samples detector in (4) involves evaluating the posterior (or some monotonically increasing function of it) for the random samples. A simpler approach is to make a decision based on the *empirical distribution* of $(X'_1, \ldots, X'_N)$ and take the *empirical mode*

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

which is the most frequent symbol among the samples. It is natural to expect that with more random samples, we would achieve a smaller error probability.

Variation 3.

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

where

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

The proof is omitted. This distribution-free result is not readily scalable with large N. With some regularity conditions on the posterior, however, we can establish the following.

Variation 4. *Let*

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

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

$$\Delta(y) \coloneqq q_1(y) - q_2(y)$$

and

$$\epsilon(\delta) = \mathsf{P}\{\Delta(Y) \leq \delta\}. \qquad (10)$$

Then, for any $\delta > 0$,

$$\begin{aligned}
\mathsf{P}\{X \neq \hat{X}_N\} \leq P_e^* + \min\{&(|\mathcal{X}| - 1)(e^{-\frac{\delta^2 N}{2}} + \epsilon(\delta)), \\
&8(N+1)(e^{-\frac{\delta^2 N}{128}} + \epsilon(\delta))\}.
\end{aligned}$$

In particular, if $\operatorname{ess\,inf}_{y \in \mathcal{Y}} \Delta(y) > 0$, $\mathsf{P}\{X \neq \hat{X}_N\}$ *converges to* P_e^* *exponentially in* N.

The proof of Variation 4 follows exactly as that of Variation 2, with the use of the following lemma.

Lemma 3. *For every* $y \in \mathcal{Y}$,

$$\begin{aligned}
\mathsf{P}\{\hat{X}_N \notin \mathcal{X}^*(y) | Y = y\} \leq \min\{&(|\mathcal{X}| - 1) e^{-\frac{\Delta^2(y) N}{2}}, \\
&8(N+1) e^{-\frac{\Delta^2(y) N}{128}}\}.
\end{aligned}$$

Proof: For simplicity of notation, we drop the dependence on y. Suppose that $k = |\mathcal{X}^*| < \infty$. Let $\pi(x|x'^N) \coloneqq \sum_{i=1}^N \mathbb{1}_{\{x'_i = x\}}$ be the number of occurrences of x in the sequence x'^N. Assume without loss of generality that $1 \in \mathcal{X}^*$. Then

$$\begin{aligned}
\mathsf{P}\{\hat{X}_N \notin \mathcal{X}^*\} &= \sum_{x \notin \mathcal{X}^*} \mathsf{P}\{\hat{X}_N = x\} \\
&\leq \sum_{x \notin \mathcal{X}^*} \mathsf{P}\{\pi(x|X'^N) \geq \pi(x'|X'^N), x' \in \mathcal{X}^*\} \\
&\leq \sum_{x \notin \mathcal{X}^*} \mathsf{P}\{\pi(x|X'^N) \geq \pi(1|X'^N)\} \\
&= \sum_{x \notin \mathcal{X}^*} \mathsf{P}\left\{\sum_{i=1}^N \mathbb{1}_{\{X'_i = x\}} - \mathbb{1}_{\{X'_i = 1\}} \geq 0\right\} \\
&= \sum_{x \notin \mathcal{X}^*} \mathsf{P}\left\{\sum_{i=1}^N Z_{x,i} \geq 0\right\},
\end{aligned}$$

where

$$Z_{x,i} = \mathbb{1}_{\{X'_i = x\}} - \mathbb{1}_{\{X'_i = 1\}} = \begin{cases} 1 & \text{w.p. } p(x), \\ -1 & \text{w.p. } p(1), \\ 0 & \text{w.p. } 1 - (p(1) + p(x)). \end{cases}$$

Clearly, $\mathsf{E}[Z_{x,i}] = p(x) - p(1) < 0$ for $x \notin \mathcal{X}^*$ by assumption. Since $(Z_{x,i})_{i=1}^N$ are i.i.d. and bounded random variables, by Hoeffding's inequality we have for each $x \notin \mathcal{X}^*$

$$\begin{aligned}&\mathsf{P}\left\{\sum_{i=1}^N Z_{x,i} \ge 0\right\}\\ &\le \mathsf{P}\left\{\sum_{i=1}^N Z_{x,i} - N(p(x)-p(1)) \ge N(p(1)-p(x))\right\}\\ &\le \exp\left(-\frac{(p(1)-p(x))^2}{2}N\right).\end{aligned}$$

Finally, by the union bound,

$$\begin{aligned}\mathsf{P}\{\hat{X}_N \notin \mathcal{X}^*\} &\le \sum_{x\notin\mathcal{X}^*} \exp\left(-\frac{(p(1)-p(x))^2}{2}N\right)\\ &\le (M-k)\exp\left(-\frac{(q_1-q_2)^2}{2}N\right).\end{aligned}$$

The case for $|\mathcal{X}| = \infty$ can be handled by the Vapnik–Chervonenkis (VC) theory [16]. First note that the event $\{\hat{X}_N \notin \mathcal{X}^*\}$ is included in the event

$$A := \left\{\sup_{x\in\mathbb{N}}\left|\frac{1}{N}\pi(x|X^N) - p(x)\right| > \frac{\Delta}{2}\right\}.$$

Also, the N-th shatter coefficient of $\{\{x\}\colon x \in \mathbb{N}\}$ becomes $N+1$ by definition. Hence, by the VC theorem, we have

$$\mathsf{P}(\hat{X}_N \notin \mathcal{X}^*) \le \mathsf{P}(A) \le 8(N+1)e^{-N\Delta^2/128},$$

which completes the proof. ∎

III. Estimation from Posterior Sampling

We now consider the problem of estimating a signal X from its noisy observation Y under a specified distortion function $d : \mathcal{X}\times\mathcal{X} \to [0,\infty)$. As in Sections I and II, we discuss the performance of the posterior sampling estimator $X' \sim f(x|y)$ compared to the optimal estimator

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

A. Absolute Loss

Suppose that the distortion function is the absolute loss, i.e., $d(x,\hat{x}) = |x-\hat{x}|$. Given $\{Y=y\}$, the optimal estimator is the *median*

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

of the posterior cumulative distribution function (cdf) $F(x|y)$. Let X' be a conditionally i.i.d. copy of X. Since the absolute loss is a metric, Lemma 1 implies the following.

Variation 5.

$$\mathsf{E}[|X-\hat{x}^*(Y)|] \le \mathsf{E}[|X-X'|] \le 2\,\mathsf{E}[|X-\hat{x}^*(Y)|]. \quad (11)$$

In words, posterior sampling has the factor-of-two performance guarantee from the MMAE estimate.

We now consider taking multiple samples $X'_1, X'_2, \dots, X'_N$ from the posterior $f(x|y)$. Define the empirical cdf

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

and take

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

to be the empirical median.

The following result compares the expected loss of this posterior sampling estimator to the MMAE.

Variation 6. *Suppose that $|X| \le B < \infty$ almost surely and that the posterior cdf $F(x|y)$ satisfies*

$$|F(\hat{x}^*(y)+\alpha|y) - 1/2| \ge L|\alpha|,$$

for every $y \in \mathcal{Y}$ and $\alpha \in (-r,r)$ for some $L, r > 0$. Then, for any $\epsilon > 0$,

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

To prove this result, we first argue the convergence in probability via the following lemma.

Lemma 4. *Under the condition of Variation 6, for any $\epsilon > 0$,*

$$\mathsf{P}\left\{|\hat{X}_N - \hat{x}^*(Y)| > \epsilon \middle| Y=y\right\} \le 2e^{-2L^2(\epsilon\wedge r)^2 N}.$$

Proof: Let $\epsilon > 0$ and define $\delta := \epsilon \wedge r$. If

$$\sup_x |F(x|y) - \hat{F}_N(x|y)| < L\delta, \quad (12)$$

then

$$F(\hat{X}_N|y) > \hat{F}_N(\hat{X}_N|y) - L\delta \ge \frac{1}{2} - L\delta \ge F(\hat{x}^*(y)-\delta|y),$$

where the last inequality is due to the assumption on $F(x|y)$. It follows that $\hat{X}_N > \hat{x}^*(y)-\delta$ since $F(x|y))$ is nondecreasing in x. Following a similar argument, (12) also implies that for any $\alpha > 0$

$$F(\hat{X}_N - \alpha) < \hat{F}_N(\hat{X}_N - \alpha) + L\delta < \frac{1}{2} + L\delta \le F(\hat{x}^*(y)+\delta)$$

and thus that $\hat{X}_N - \alpha < \hat{x}^*(y) + \delta$. Letting $\alpha \to 0$ yields $\hat{X}_N \le \hat{x}^*(y) + \delta$. Now by the Dvoretzky–Kiefer–Wolfowitz inequality [17],

$$\begin{aligned}&\mathsf{P}\left\{|\hat{X}_N - \hat{x}^*(Y)| > \epsilon \middle| Y=y\right\}\\ &\le \mathsf{P}\left\{\sup_x |F(x|Y) - \hat{F}_N(x|Y)| > L\delta \middle| Y=y\right\}\\ &\le 2e^{-2L^2\delta^2 N},\end{aligned}$$

which completes the proof. ∎

We are now ready to prove Variation 6.

Proof of Variation 6: Let $\epsilon > 0$. By the triangle inequality, we have

$$\begin{aligned}&\mathsf{E}\left[|X-\hat{X}_N| \middle| Y=y\right]\\ &\le \mathsf{E}\left[|X-\hat{x}^*(Y)| \middle| Y=y\right] + \mathsf{E}\left[|\hat{X}_N - \hat{x}^*(Y)| \middle| Y=y\right].\end{aligned}$$

We now bound the second term as

$$\begin{aligned}
&\mathsf{E}\left[|\hat{X}_N - \hat{x}^*(Y)|\middle|Y=y\right] \\
&= \mathsf{E}\left[|\hat{X}_N - \hat{x}^*(y)|\mathbb{1}_{\{|\hat{X}_N-\hat{x}^*(y)|\le\epsilon\}}\middle|Y=y\right] \\
&\qquad + \mathsf{E}\left[|\hat{X}_N - \hat{x}^*(y)|\mathbb{1}_{\{|\hat{X}_N-\hat{x}^*(y)|>\epsilon\}}\middle|Y=y\right] \\
&\le \epsilon + \mathsf{E}\left[2B\,\mathbb{1}_{\{|\hat{X}_N-\hat{x}^*(y)|>\epsilon\}}\middle|Y=y\right] \qquad (13)\\
&\le \epsilon + 2B\,\mathsf{P}\left\{|\hat{X}_N - \hat{x}^*(Y)| > \epsilon\middle|Y=y\right\} \\
&\le \epsilon + 4Be^{-2L^2(\epsilon\wedge r)^2 N}, \qquad (14)
\end{aligned}$$

where (13) follows since $|\hat{X}_N - \hat{x}^*(y)| \le |\hat{X}_N| + |\hat{x}^*(y)| \le 2B$ by the triangle inequality and by the assumption that $|X| \le B$ almost surely, and (14) follows by Lemma 4. Taking expectation w.r.t. Y completes the proof. ∎

B. Quadratic Loss

Suppose that the distortion function is the quadratic loss $d(x,\hat{x}) = (x-\hat{x})^2$. The optimal estimator in this case is the conditional expectation (mean)

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

Note that the quadratic loss is not a metric (it does not follow the triangle inequality) and hence Lemma 1 is not applicable here. Nonetheless, we have

$$\begin{aligned}
&\mathsf{E}[(X-X')^2|Y=y] \\
&= \mathsf{E}[(X-\hat{x}^*(y)-X'+\hat{x}^*(y))^2|Y=y] \\
&= \mathsf{E}[(X-\hat{x}^*(y))^2|Y=y] + \mathsf{E}[(X'-\hat{x}^*(y))^2|Y=y],
\end{aligned}$$

which implies the following factor-of-two result.

Variation 7.

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

We now consider taking multiple samples from the posterior $f(x|y)$. Define our posterior sampling estimator as the empirical mean

$$\hat{X}_N = \frac{1}{N}(X'_1 + \ldots + X'_N).$$

Unlike detection or MMAE estimation, the convergence in this case linear rather than exponential.

Variation 8.

$$\mathsf{E}[(X-\hat{X}_N)^2] = \left(1+\frac{1}{N}\right)\mathsf{E}[(X-\hat{x}^*(Y))^2]. \qquad (15)$$

We can also characterize the performance of the posterior sampling estimator through its convergence to X in probability.

Variation 9. *Suppose that $|X| \le B < \infty$ almost surely. Then, for any $\epsilon > 0$ and $0 < \delta < 1$*

$$\mathsf{P}\{|X-\hat{X}_N| \ge \epsilon\} \le \mathsf{P}\{|X-\hat{x}^*(Y)| \ge \bar{\delta}\epsilon\} + 2e^{-\delta^2\epsilon^2 N/4B^2},$$

where $\bar{\delta} := 1-\delta$.

Proof: Let $\epsilon > 0$ and $\delta \in (0,1)$. For a metric d over an alphabet $\mathcal{X}$, and for any random variables X, X', and X'' over $\mathcal{X}$, by the triangle inequality and union bound, we have

$$\mathsf{P}\{d(X,X') \ge \epsilon\} \le \mathsf{P}\{d(X,\tilde{X}) \ge \bar{\delta}\epsilon\} + \mathsf{P}\{d(X',\tilde{X}) \ge \delta\epsilon\}.$$

Using this fact conditioned on $\{Y=y\}$, and replacing $X' = \hat{X}_N$ and $X'' = \hat{x}^*(Y)$, we have

$$\begin{aligned}
&\mathsf{P}\left\{|X-\hat{X}_N| \ge \epsilon\middle|Y=y\right\} \\
&\quad\le \mathsf{P}\left\{|X-\hat{x}^*(Y)| \ge \bar{\delta}\epsilon\middle|Y=y\right\} \\
&\qquad + \mathsf{P}\left\{|\hat{X}_N-\hat{x}^*(Y)| \ge \delta\epsilon\middle|Y=y\right\} \\
&\quad\le \mathsf{P}\left\{|X-\hat{x}^*(Y)| \ge \bar{\delta}\epsilon\middle|Y=y\right\} + 2e^{-\delta^2\epsilon^2 N/4B^2}
\end{aligned}$$

where the last inequality follows by Hoeffding's inequality since $\mathsf{E}[\hat{X}_N|Y=y] = \hat{x}^*(y)$. The desired claim follows by averaging w.r.t. Y. ∎

REFERENCES

[1] J. Liu, P. Cuff, and S. Verdú, "On α-decodability and α-likelihood decoder," in *Proc. 55th Ann. Allerton Conf. Comm. Control Comput.*, Monticello, IL, Oct. 2017.

[2] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[3] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2201–2214, 2002.

[4] B. Hassibi and H. Vikalo, "On the sphere-decoding algorithm—I. Expected complexity," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2806–2818, 2005.

[5] H. Vikalo and B. Hassibi, "On the sphere-decoding algorithm—II. Generalizations, second-order statistics, and applications to communications," *IEEE Trans. Signal Process.*, vol. 53, pp. 2819–2834, 2005.

[6] H. Zhu, B. Farhang-Boroujeny, and R. R. Chen, "On performance of sphere decoding and Markov chain Monte Carlo detection methods," in *IEEE 6th Workshop on Signal Processing Advances in Wireless Communications*, June 2005, pp. 86–90.

[7] B. Hassibi, M. Hansen, A. G. Dimakis, H. A. J. Alshamary, and W. Xu, "Optimized Markov chain Monte Carlo for signal detection in MIMO systems: An analysis of the stationary distribution and mixing time," *IEEE Trans. Signal Process.*, vol. 62, no. 17, pp. 4436–4450, Sept 2014.

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

[9] E. C. Song, P. Cuff, and H. V. Poor, "The likelihood encoder for lossy compression," *IEEE Trans. Inf. Theory*, vol. 62, pp. 1836–1849, 2016.

[10] J. Scarlett, A. Martinez, and A. G. i Fàbregas, "The likelihood decoder: error exponents and mismatch," in *Proc. IEEE Int. Symp. Inf. Theory*, Hong Kong, Jul. 2015, pp. 86–90.

[11] N. Merhav, "The generalized stochastic likelihood decoder: Random coding and expurgated bounds," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 5039–5051, Aug. 2017.

[12] R. M. Neal, "Monte Carlo decoding of LDPC codes," Dept. of Computer Science, University of Toronto, Tech. Rep., 2001, presented at *ICTP Works. Statis. Phys. Capacity-Approaching Codes*.

[13] M. Mezard and A. Montanari, *Information, Physics, and Computation*. Oxford University Press, 2009.

[14] A. Bhatt, J.-T. Huang, Y.-H. Kim, J. Ryu, and P. Sen, "Monte Carlo methods for randomized likelihood decoding," in *Proc. 56th Ann. Allerton Conf. Comm. Control Comput.*, Monticello, IL, Oct. 2018.

[15] S. Beigi and A. Gohari, "Quantum achievability proof via collision relative entropy," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7980–7986, 2014.

[16] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. Appl.*, vol. 16, no. 2, pp. 264–280, 1971.

[17] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator," *Ann. Math. Stat.*, vol. 27, no. 3, pp. 642–669, 1956.