Learning Joint Quantizers for Reconstruction and Prediction

Maxim Raginsky

***Abstract*—We consider the problem of empirical design of variable-rate quantizers for reconstruction and prediction. When a discriminative model (conditional distribution of the unobserved output given the observed input) is known or can be accurately estimated from a separate training set, we show that this problem reduces to designing a certain type of a generalized quantizer by means of empirical risk minimization on *unlabeled* input samples only. We derive a high-probability upper bound on the resulting expected performance of such a quantizer in terms of the training sample size and the complexity parameters of the reconstruction and the prediction problems. We also discuss two illustrative examples: binary classification with absolute loss and the information bottleneck.**

I. Introduction

The problem of designing compressed representations of stochastic data sources subject to a reconstruction fidelity criterion is well-understood theoretically [1], [2]. There is also a wide variety of practical methods and algorithms which, though not necessarily optimal from a theoretical point of view, have been developed based on extensive domain knowledge and, show good empirical performance on many sources of interest (such as audio, images or video) [3]. On the other hand, the problem of designing compressed representations for inference or prediction is still lacking a theoretical framework comparable to that for lossy source coding. The main challenge is that there seems to be no universal way of quantifying the amount of "predictively relevant" information, although several proposals based either on decision-theoretic or on information-theoretic ideas have been put forward [4]–[8]. Moreover, one can easily come up with examples where the amount of information needed for reconstruction is much greater than what would be required for prediction, or vice versa.

It is often the case in practice that there is no readily available probabilistic model for the data source of interest. In such situations, one resorts to *empirical design* based on training data sampled from the source, and techniques from statistical learning theory can be used to analyze consistency and convergence to optimal performance (see [9] for applications to lossy source coding). In this paper, we apply learning-theoretic ideas to the problem of designing compressed representations that attempt to preserve enough information to enable both good reconstruction and good prediction. We show that this problem can be reduced to designing a generalized quantizer with a modified fidelity criterion that incorporates both the reconstruction error term and the prediction error term, as well as the description length in bits. Our main result (Theorem 1) is a high-probability upper bound on the gap between the expected performance of an empirically learned generalized quantizer and the theoretical optimum. We also discuss two illustrative examples: binary classification with absolute loss and the information bottleneck.

The author is with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois, Urbana, IL 61801, USA. E-mail: maxim@illinois.edu.

Research supported in part by DARPA under grant no. N66001-13-1-4004 and in part by NSF under CAREER award no. CCF-1254041.

II. Preliminaries

We start by fixing some useful terminology and notation. We assume in the sequel that all spaces are standard Borel [10] (so that, in particular, all conditional distributions have regular versions), and that all maps under consideration are measurable with respect to appropriate σ-algebras.

A. Quantization

A variable-rate quantizer with input space X and reproduction space Z is a pair $q = (\varphi, \psi)$, where $\varphi : \mathsf{X} \to \mathsf{S}$ is the *encoder*, $\psi : \mathsf{S} \to \mathsf{Z}$ is the *decoder*, and $\mathsf{S} \subseteq \{0,1\}^*$ is a countable collection of binary strings satisfying Kraft's inequality

$$\sum_{s \in \mathsf{S}} 2^{-\operatorname{len}(s)} \leq 1, \tag{1}$$

where $\operatorname{len}(s)$ denotes the length of $s \in \{0,1\}^*$ in bits. The quantizer q maps each point $x \in \mathsf{X}$ to a reproduction $z = \psi \circ \varphi(x) \in \mathsf{Z}$ (with a slight abuse of notation, we will also write $z = q(x)$). We are given a *distortion function* $d : \mathsf{X} \times \mathsf{Z} \to \mathbb{R}^+$, so that $d(x,z) = d\big(x, q(x)\big)$ is the distortion incurred in representing x by its quantized version z. Also, $\operatorname{len}\big(\varphi(x)\big)$ is the *description length* of x. If q is fed with a random input X according to a probability law P, then the *expected distortion* and the *rate* of q are given by

$$D_P(q) \triangleq \mathbb{E}[d(X, q(X))] \text{ and } R_P(q) \triangleq \mathbb{E}[\operatorname{len}(\varphi(X))]. \tag{2}$$

It is convenient [9], [11] to absorb the distortion and the rate into a single performance measure, the *Lagrangian*

$$L_P(q,\lambda) \triangleq D_P(q) + \lambda R_P(q) \tag{3}$$

$$= \mathbb{E}\left[d(X, q(X)) + \lambda \cdot \operatorname{len}(\varphi(X))\right], \tag{4}$$

where $\lambda \geq 0$ is the Lagrange multiplier that controls the distortion-rate trade-off. Given a collection of quantizers $\mathcal{Q}$, the optimum Lagrangian performance on P is given by

$$L_P^*(\mathcal{Q}, \lambda) \triangleq \inf_{q \in \mathcal{Q}} L_P(q, \lambda). \tag{5}$$

It is easy to show that, for any quantizer $q = (\varphi, \psi)$ and a Lagrange multiplier λ, we can construct another quantizer $q' = (\varphi', \psi')$ with $\psi' = \psi$ and

$$\varphi'(x) \triangleq \arg\min_{s \in \mathsf{S}} \{d(x, \psi(s)) + \lambda \cdot \mathrm{len}(s)\}, \tag{6}$$

so that $L_P(q') \leq L_P(q)$. Thus, we may always assume that our quantizers have (modified) nearest-neighbor encoders.

B. Prediction

In a statistical prediction problem, we have a random couple $(X, Y) \in \mathsf{X} \times \mathsf{Y}$ with joint probability law $\mathbb{P}$, where X is the input (or feature) and Y is the output (or label). A *predictor* is a mapping $f : \mathsf{X} \to \mathsf{U}$, where U is some *action space*. Thus, each input point $x \in \mathsf{X}$ is mapped to an action $u = f(x)$. We are given a *loss function* $\ell : \mathsf{X} \times \mathsf{Y} \times \mathsf{U} \to \mathbb{R}^+$, so that $\ell(x, y, u) = \ell\big(x, y, f(x)\big)$ is the loss incurred when taking action u in response to the input-output pair (x, y) (note that the action u is allowed to depend only on the input x, so effectively we need to *predict* the corresponding input y without seeing it in order to settle on a good action, hence the term "prediction"). The *expected loss* of f is given by

$$\Lambda_{\mathbb{P}}(f) \triangleq \mathbb{E}\left[\ell\left(X, Y, f(X)\right)\right]. \tag{7}$$

III. Problem formulation

We are interested in the following problem that combines the reconstruction and the prediction aspects: We have a random couple $(X, Y) \in \mathsf{X} \times \mathsf{Y}$, an action space U, a collection $\mathcal{Q}$ of quantizers with input and reproduction spaces both equal to X, and a collection $\mathcal{F}$ of predictors $f : \mathsf{X} \to \mathsf{U}$. Given a quantizer $q = (\varphi, \psi) \in \mathcal{Q}$ and a predictor $f \in \mathcal{F}$, we observe the input X, quantize it to $\widehat{X} = q(X) = \psi \circ \varphi(X)$, and then take an action $U = f(\widehat{X}) = f \circ q(X)$. Our goal is to jointly choose q and f to simultaneously guarantee small values of expected distortion, rate, and prediction loss. Once again, we adopt the Lagrangian viewpoint and quantify the performance of a pair $(q, f) \in \mathcal{Q} \times \mathcal{F}$ by the scalar quantity

$$\mathcal{L}_{\mathbb{P}}(q, f, \lambda, \mu) \triangleq L_{\mathbb{P}_X}(q, \lambda) + \mu \cdot \Lambda_{\mathbb{P}}(f \circ q) \tag{8}$$

$$\equiv \mathbb{E}\left[d(X, q(X)) + \lambda \cdot \mathrm{len}(\varphi(X)) + \mu \cdot \ell(X, Y, f \circ q(X))\right], \tag{9}$$

where $d : \mathsf{X} \times \mathsf{X} \to \mathbb{R}^+$ is the reconstruction distortion function, $\ell : \mathsf{X} \times \mathsf{Y} \times \mathsf{U} \to \mathbb{R}^+$ is the prediction loss function, and $\lambda, \mu \geq 0$ are the Lagrange multipliers. The optimal performance achievable by $\mathcal{Q} \times \mathcal{F}$ on $\mathbb{P}$ is

$$\mathcal{L}^*_{\mathbb{P}}(\mathcal{Q}, \mathcal{F}, \lambda, \mu) \triangleq \inf_{q \in \mathcal{Q}} \inf_{f \in \mathcal{F}} \mathcal{L}_{\mathbb{P}}(q, f, \lambda, \mu). \tag{10}$$

A. Reduction to generalized quantization

We now reduce the above problem to a quantization problem with input space X and reproduction space $\mathsf{Z} = \mathsf{X} \times \mathsf{U}$. Let $\mathbb{P}_{Y|X}$ denote the regular conditional distribution of Y given X, and define a new loss function $\widetilde{\ell} : \mathsf{X} \times \mathsf{U} \to \mathbb{R}^+$ by

$$\widetilde{\ell}(x, u) \triangleq \mathbb{E}[\ell(X, Y, u) | X = x] \tag{11}$$

$$\equiv \int_{\mathsf{Y}} \mathbb{P}_{Y|X}(\mathrm{d}y|x)\ell(x, y, u) \tag{12}$$

(this is simply the conditional expectation of the loss $\ell(X, Y, u)$ w.r.t. the "discriminative model" $\mathbb{P}_{Y|X=x}$). Then for any q and f we can write

$$\mathcal{L}_{\mathbb{P}}(q, f, \lambda, \mu)$$

$$= \mathbb{E}[d(X, q(X)) + \lambda \cdot \mathrm{len}(\varphi(X))] + \mu \cdot \mathbb{E}\Big[\mathbb{E}\big[\ell(X, Y, f \circ q(X))\big|X\big]\Big] \tag{13}$$

$$= \mathbb{E}\left[d(X, q(X)) + \mu \cdot \widetilde{\ell}(X, f \circ q(X))\right] + \lambda \cdot \mathbb{E}[\mathrm{len}(\varphi(X))]. \tag{14}$$

For any $q = (\varphi, \psi) \in \mathcal{Q}$ and $f \in \mathcal{F}$, we can define a new quantizer $\widetilde{q} = (\widetilde{\varphi}, \psi)$ with

- input space X,
- reproduction space $\mathsf{Z} = \mathsf{X} \times \mathsf{U}$,
- encoder $\widetilde{\varphi} = \varphi$, and
- decoder $\widetilde{\psi} : \mathsf{S} \to \mathsf{Z}$ with $\widetilde{\psi}(s) = (\psi(s), f \circ \psi(s))$.

If we also define a distortion function $d_\mu : \mathsf{X} \times \mathsf{Z} \to \mathbb{R}^+$ by

$$d_\mu(x, z) = d_\mu(x, (\widehat{x}, u)) \triangleq d(x, \widehat{x}) + \mu \cdot \widetilde{\ell}(x, u), \tag{15}$$

we can recognize the expression in (14) as the Lagrangian

$$L_{\mathbb{P}_X}(\widetilde{q}, \lambda; \mu) \triangleq \mathbb{E}\left[d_\mu(X, \widetilde{q}(X))\right] + \lambda \cdot \mathbb{E}[\mathrm{len}(\widetilde{\varphi}(X))] \tag{16}$$

of the *generalized quantizer* $\widetilde{q}(X) = \big(q(X), f \circ q(X)\big)$.

In a nutshell, if we let $\widetilde{\mathcal{Q}}$ denote the collection of all generalized quantizers $\widetilde{q}$ of the above form, then we have

$$\mathcal{L}^*_{\mathbb{P}}(\mathcal{Q}, \mathcal{F}, \lambda, \mu) = L^*_{\mathbb{P}_X}(\widetilde{\mathcal{Q}}, \lambda; \mu). \tag{17}$$

Note that we have effectively removed the label Y from the picture: it enters indirectly, through the "posterior expected loss" $\widetilde{\ell}(x, u) = \mathbb{E}[\ell(X, Y, u)|X = x]$. This reduction is conceptually similar to indirect rate-distortion problems (see [12] and references therein).

IV. Empirical quantizer design

In this paper, we are interested in the situation where the joint distribution $\mathbb{P}$ is unknown, and we must instead *learn* a suitable quantizer-predictor pair on the basis of a training sequence $\{(X_i, Y_i)\}_{i=1}^m$ of i.i.d. samples from $\mathbb{P}$. Of course, when we do not know $\mathbb{P}$, the above reduction to generalized quantization does not hold – indeed, we cannot compute $\widetilde{\ell}(\cdot, \cdot)$ exactly since we do not know $\mathbb{P}_{Y|X}$. However, we may first learn a good discriminative model from a large labeled sample, and then use this learned model in a plug-in fashion on another sample when designing our quantizer and predictor. Moreover, once we have obtained such a discriminative model, we no longer need labeled data – our problem reduces to empirical quantizer design (albeit with an augmented reproduction space), which can be done only with unlabeled input samples.

This may seem like a roundabout way of learning a quantizer-predictor pair: first we learn a very good randomized predictor, and then we effectively forget (or marginalize away) what we have learned in order to design a quantizer and an inferior predictor. However, this separation-based approach makes sense as soon as we consider the case when the

feature X is first encoded into a binary representation at some remote observation station, and then this binary representation is transmitted to a decision-maker who will use it to both reconstruct the feature and to take a decision based on the reconstruction. In this setting, the encoder has strictly better information than the decision-maker, which is actually helpful, since we can guarantee that the quantized representation of the feature retains as much of this information as possible for a given choice of the Lagrange multipliers.

From now on we will assume that the conditional distribution $\mathbb{P}_{Y|X}$ is known (at least to the extent needed to compute $\widetilde{\ell}$), and focus on generalized quantizer design with *unlabeled* input samples. Given n i.i.d. samples $X_1, \ldots, X_n$ from the marginal distribution $\mathbb{P}_X$ of the input variable (or feature) X, we seek a solution to the following Empirical Risk Minimization (ERM) problem:

$$\widetilde{q}_n \equiv (\widehat{q}_n, \widehat{f}_n) = \arg\min_{\widetilde{q} \in \widetilde{\mathcal{Q}}} \widehat{L}_n(\widetilde{q}, \lambda; \mu), \tag{18}$$

where we have defined the empirical Lagrangian

$$\widehat{L}_n(\widetilde{q}, \lambda; \mu) = \frac{1}{n} \sum_{i=1}^{n} \left\{ d_\mu\left(X_i, \widetilde{q}(X_i)\right) + \lambda \cdot \mathrm{len}(\widetilde{\varphi}(X_i)) \right\} \tag{19}$$

for any quantizer $\widetilde{q} = (q, f \circ q)$.

V. The main result and some examples

Our main result gives a high-probability bound on the generalization performance of the ERM solution (18):

Theorem 1. *Suppose that both the reconstruction distortion d and the prediction loss ℓ are uniformly bounded:*

$$d_{\max} \triangleq \sup_{x \in \mathsf{X}} \sup_{\widehat{x} \in \mathsf{X}} d(x, \widehat{x}) < +\infty \tag{20}$$

$$\ell_{\max} \triangleq \sup_{x \in \mathsf{X}} \sup_{y \in \mathsf{Y}} \sup_{u \in \mathsf{U}} \ell(x, y, u) < +\infty. \tag{21}$$

Let $\mathcal{A}$ denote the collection of all sets of the form

$$\left\{ x \in \mathsf{X} : d(x, \widehat{x}) + \mu \widetilde{\ell}(x, u) \le r \right\} \tag{22}$$

for all $\widehat{x} \in \mathsf{X}$, $u \in \mathsf{U}$, $r > 0$, and suppose that $\mathcal{A}$ is a Vapnik–Chervonenkis (VC) class with VC dimension V [13]. Let $\mathcal{Q}$ be the collection of all quantizers on X, and let $\mathcal{F}$ be the collection of all mappings $f : \mathsf{X} \to \mathsf{U}$. Then there exists some absolute constant $C > 0$, such that, for any $\lambda > 0$ and $\mu \ge 0$,

$$\mathcal{L}_{\mathbb{P}}(\widehat{q}_n, \widehat{f}_n, \lambda, \mu) \le \mathcal{L}^*_{\mathbb{P}}(\mathcal{Q}, \mathcal{F}, \lambda, \mu) + T\left(C\sqrt{\frac{VN \log N}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}} \right) \tag{23}$$

with probability at least $1 - \delta$, where

$$T = 3\left(d_{\max} + \mu \ell_{\max}\right) \quad \text{and} \quad N = 2^{2T/3\lambda}. \tag{24}$$

Remark 1. When $\ell \equiv 0$, we recover existing finite-sample bounds for empirically optimal variable-rate quantizers [9].

Remark 2. As we had mentioned earlier, the discriminative model $\mathbb{P}_{Y|X}$ must be learned separately. For instance, suppose we use an independent labeled sample of size m to construct an estimate $\widetilde{\ell}_m$ of $\widetilde{\ell}$, such that $\left\| \widetilde{\ell} - \widetilde{\ell}_m \right\|_\infty \le \varepsilon_m$ with probability at least $1 - \delta$. Then we can plug this estimate into the ERM procedure (18) as a proxy for $\widetilde{\ell}$. As a consequence, the error bound of Theorem 1 will pick up an extra term $\mu\varepsilon_m$, and the new bound will hold with probability at least $1 - 2\delta$.

The proof of the theorem is given in the next section. Here, we discuss two specific examples.

A. Binary classification with absolute loss

Let X be a compact subset of $\mathbb{R}^d$, $\mathsf{Y} = \{0, 1\}$, and $\mathsf{U} = [0, 1]$. The reconstruction fidelity is measured by the squared Euclidean distance, $d(x, \widehat{x}) = \|x - \widehat{x}\|^2$, whereas the prediction error is measured by the absolute loss $\ell(x, y, u) = |y - u|$ (notice that the loss does not directly depend on the feature x). The boundedness conditions of Theorem 1 are automatically satisfied. Defining the usual regression function $\eta(x) \triangleq \mathbb{E}[Y|X = x] \equiv \mathbb{P}(Y = 1|X = x)$ [13], we can write

$$\widetilde{\ell}(x, u) = u + (1 - 2u)\eta(x). \tag{25}$$

Thus, we are interested in bounding the VC dimension of the collection $\mathcal{A}$ of all sets of the form

$$\left\{ x \in \mathsf{X} : \|x - \widehat{x}\|^2 + \mu\left(u + (1 - 2u)\eta(x)\right) \le r \right\} \tag{26}$$

for all $\widehat{x} \in \mathsf{X}$, $u \in [0, 1]$, $r > 0$. The following lemma follows from a well-known result that an m-dimensional space of real-valued functions is a VC-subgraph class with VC dimension $\le m$ [13, Theorem 13.9]:

Lemma 1. *Suppose that η can be represented as a linear combination $\sum_{j=1}^{k} \alpha_j \xi_j$ of k linearly independent functions $\xi_1, \ldots, \xi_k$ on X. Then $V(\mathcal{A}) \le d + 2k + 3$.*

B. Information bottleneck

Again, let X be a compact subset of $\mathbb{R}^d$, let Y be arbitrary, and suppose that all conditional distributions $\mathbb{P}_{Y|X=x}$, $x \in \mathsf{X}$, are dominated by a given σ-finite measure ν on Y. We again take the squared Euclidean distance $d(x, \widehat{x}) = \|x - \widehat{x}\|^2$ as our reconstruction criterion. As our action space U we take a suitable subset of the space of all ν-densities $u : \mathsf{Y} \to \mathbb{R}^+$ with $\int u \, \mathrm{d}\nu = 1$, and consider the prediction loss of the form

$$\ell(x, y, u) = \phi\left(\frac{u(y)}{p_x(y)}\right), \tag{27}$$

where ϕ is a convex function on $\mathbb{R}^+$ and where $p_x \triangleq \mathrm{d}\mathbb{P}_{Y|X=x}/\mathrm{d}\nu$. Then

$$\widetilde{\ell}(x, u) = \int_{\mathsf{Y}} \nu(\mathrm{d}y) p_x(y) \phi\left(\frac{u(y)}{p_x(y)}\right) \equiv D_\phi(u \| p_x), \tag{28}$$

which we recognize as the ϕ-divergence between the densities u and p_x [14], [15]. For example, if we let $\phi(t) = t \log t$, then we get the Kullback–Leibler divergence

$$\widetilde{\ell}(x, u) = D(p_x \| u); \tag{29}$$

with $\phi(t) = (1-\sqrt{t})^2$, we get the squared Hellinger distance

$$\widetilde{\ell}(x,u) = \int_{\mathsf{Y}} \nu(\mathrm{d}y) \left(\sqrt{p_x(y)} - \sqrt{u(y)}\right)^2 = H^2(p_x, u). \tag{30}$$

Loss functions of this type underlie the so-called information bottleneck method [4], [6] for constructing compressed representations of stochastic data sources that retain as much "predictively relevant" information as possible (see also [8] for a different application of ϕ-divergences in the context of surrogate loss functions for binary classification). In fact, if we ignore reconstruction entirely by letting $d(\cdot,\cdot) = 0$ and set $\lambda \equiv 1$, then we end up with an optimization problem of the form

$$\inf_{q,f} \mathbb{E}\left\{\mathrm{len}\left(\varphi(X)\right) + \mu \cdot D_\phi\big(f_{q(X)} \big\| p_X\big)\right\}, \tag{31}$$

where, for notational consistency, we wrote $f_{q(x)}$ for the image of $q(x)$ under $f : \mathsf{X} \to \mathsf{U}$. The Shannon-theoretic counterpart of this operational problem,

$$\inf_{U} \left\{I(X;U) + \mu\, \mathbb{E}\left[D(U\|p_X)\right]\right\}, \tag{32}$$

where I is the mutual information [16] and the infimum is over all U-valued random variables (random densities) jointly distributed with X, yields the original information bottleneck formulation. By including a data reconstruction fidelity (distortion) term, we obtain a generalization of information bottleneck that can be used to design optimal compressed representations for joint reconstruction and prediction. (See also [7] for related results on quantizers with a fixed finite number of levels.)

The boundedness condition on d is satisfied by compactness of X, whereas the boundedness of ℓ can be ensured via appropriate restrictions on the likelihood ratios u/p_x and the function ϕ. We seek an upper bound on the VC dimension of the collection $\mathcal{A}$ of all sets of the form

$$\left\{x \in \mathsf{X} : \|x - \widehat{x}\|^2 + \mu D_\phi(u\|p_x) \le r\right\} \tag{33}$$

for all $\widehat{x} \in \mathsf{X}$, $u \in \mathsf{U}$, $r > 0$. The following lemma can be proved using similar methods as Lemma 1:

Lemma 2. *Let U be chosen in such a way that one can find k linearly independent functions $\xi_1, \ldots, \xi_k$ on X and a sequence $\Phi_1, \Phi_2, \ldots$ of functions in $L^1(\nu)$ with the following property: any function of the form $h(x,y) = p_x(y)\phi\left(u(y)/p_x(y)\right)$ for some $u \in \mathsf{U}$ admits a bilinear representation*

$$h(x,y) = \sum_{j=1}^{k} \sum_{m=1}^{\infty} c_{jm} \theta_j(x) \Phi_m(y),$$

such that the sequence $\left\{c_{jm}\|\Phi_m\|_{L^1(\nu)}\right\}_{m=1}^{\infty}$ is summable for every j. Then $V(\mathcal{A}) \le d + k + 2$.

VI. The proof of Theorem 1

We will need the following lemma, similar in spirit to Lemma A.3 in [17] (see also Lemma 10 in [9]):

Lemma 3. *Given two positive integers N, L, let $\widetilde{\mathcal{Q}}(N,L) \subseteq \widetilde{\mathcal{Q}}$ denote the collection of all quantizers $\widetilde{q} = (\widetilde{\varphi}, \widetilde{\psi})$ with modified nearest-neighbor encoders*

$$\widetilde{\varphi}(x) = \arg\min_{s \in \mathsf{S}} \left\{d_\mu(x, \widetilde{\psi}(x)) + \lambda \cdot \mathrm{len}(s)\right\}, \tag{34}$$

such that $|\mathsf{S}| \le N$ and $\mathrm{len}(s) \le L$ for all $s \in \mathsf{S}$. Then for any probability distribution P on X we have

$$L_P^*(\widetilde{\mathcal{Q}}, \lambda; \mu) = L_P^*\left(\widetilde{\mathcal{Q}}(N,L), \lambda; \mu\right) \tag{35}$$

with

$$\log_2 N = L = \frac{2(d_{\max} + \mu\lambda_{\max})}{\lambda}. \tag{36}$$

Proof: Given P, let $\widetilde{q}^*$ with encoder $\widetilde{\varphi}^* : \mathsf{X} \to \mathsf{S}$ and decoder $\widetilde{\psi}^* : \mathsf{S} \to \mathsf{X} \times \mathsf{U}$ achieve (or come arbitrarily close to) the optimum performance on the left-hand side of (35). Let s_0 denote the shortest binary string in S, i.e., $\mathrm{len}(s_0) = \min_{s \in \mathsf{S}} \mathrm{len}(s)$. Since $\widetilde{\varphi}^*$ is a modified nearest-neighbor encoder, for any $s \in \mathsf{S}$ and any $x \in \mathsf{X}$ such that $s = \widetilde{\varphi}^*(x)$ we have

$$d_\mu(x, \widetilde{\psi}^*(s)) + \lambda \cdot \mathrm{len}(s) \le d_\mu(x, \widetilde{\psi}^*(s_0)) + \lambda \cdot \mathrm{len}(s_0). \tag{37}$$

Thus, for any $s \in \mathsf{S}$,

$$\mathrm{len}(s) \le \frac{d_{\max} + \mu\ell_{\max}}{\lambda} + \mathrm{len}(s_0). \tag{38}$$

On the other hand, consider a (very) suboptimal quantizer $\widetilde{q}_0$ with encoder $\widetilde{\varphi}_0(x) = \mathsf{e}$ (the empty binary string of length 0) and decoder $\widetilde{\psi}_0(\mathsf{e}) = \widehat{x}_0 \times u_0$ for some arbitrary choice of $\widehat{x}_0 \in \mathsf{X}$ and $u_0 \in \mathsf{U}$. Then $\widetilde{q}_0 \in \widetilde{\mathcal{Q}}$, and

$$\lambda \cdot \mathrm{len}(s_0) \le L_P^*(\widetilde{\mathcal{Q}}, \lambda; \mu) \le L_P(\widetilde{q}_0, \lambda; \mu) \le d_{\max} + \mu\ell_{\max}. \tag{39}$$

Consequently, for any $s \in \mathsf{S}$ we have

$$\mathrm{len}(s) \le \frac{2(d_{\max} + \mu\ell_{\max})}{\lambda}. \tag{40}$$

Since the binary strings in S must satisfy Kraft's inequality, we have

$$1 \ge \sum_{s \in \mathsf{S}} 2^{-\mathrm{len}(s)} \ge |\mathsf{S}| 2^{-2(d_{\max} + \mu\ell_{\max})/\lambda}. \tag{41}$$

Thus, the optimal quantizer $\widetilde{q}^*$ has no more than $N = 2^{2(d_{\max} + \mu\ell_{\max})/\lambda}$ levels, and the description length of any $x \in \mathsf{X}$ does not exceed $L = 2(d_{\max} + \mu\ell_{\max})/\lambda$. ∎

In view of Lemma 3, both the empirically optimal quantizer $\widetilde{q}_n$ and the theoretically optimal quantizer $\widetilde{q}^*$ (i.e., the one that achieves $L^*_{\mathbb{P}_X}(\lambda;\mu)$) belong to the class $\widetilde{\mathcal{Q}}(N,L)$ with N and L given in (36). That is,

$$\widehat{L}_n(\widetilde{q}_n, \lambda; \mu) = \inf_{\widetilde{q} \in \widetilde{\mathcal{Q}}(N,L)} L_{\widehat{P}_n}(\widetilde{q}, \lambda; \mu), \tag{42}$$

$$L_{\mathbb{P}_X}(\widetilde{q}^*, \lambda; \mu) = \inf_{\widetilde{q} \in \widetilde{\mathcal{Q}}(N,L)} L_{\mathbb{P}_X}(\widetilde{q}, \lambda; \mu), \tag{43}$$

where $\widehat{P}_n = (1/n)\sum_{i=1}^n \delta_{X_i}$ is the empirical distribution induced by the sample X^n. By a standard argument we obtain

$$L_{\mathbb{P}_X}(\widetilde{q}_n,\lambda;\mu) - L_{\mathbb{P}_X}(\widetilde{q}^*,\lambda;\mu) \leq 2 \sup_{\widetilde{q}\in\widetilde{\mathcal{Q}}(N,L)} \left| L_{\widehat{P}_n}(\widetilde{q},\lambda;\mu) - L_{\mathbb{P}_X}(\widetilde{q},\lambda;\mu)\right|. \quad (44)$$

Since all quantizers under consideration have modified nearest-neighbor encoders, for any distribution P on X and any $\widetilde{q} = (\widetilde{\varphi},\widetilde{\psi})$ we have $L_P(\widetilde{q},\lambda;\mu) = \mathbb{E}_P\left[g_{\widetilde{q}}(X)\right]$, where

$$g_{\widetilde{q}}(x) \triangleq \min_{s\in\mathsf{S}} \left\{ d_\mu(x,\widetilde{\psi}(s)) + \lambda\cdot\mathrm{len}(s)\right\}. \quad (45)$$

Let $\mathcal{G}$ denote the class of all functions $g_{\widetilde{q}}$ as $\widetilde{q}$ ranges over $\widetilde{\mathcal{Q}}(N,L)$. Denoting the supremum on the right-hand side of (44) by Δ_n, we can write

$$\Delta_n = \sup_{g\in\mathcal{G}} \left| \mathbb{E}_{\widehat{P}_n} g - \mathbb{E}_{\mathbb{P}_X} g\right|. \quad (46)$$

Any g in $\mathcal{G}$ is bounded between 0 and $T \triangleq 3\left(d_{\max} + \mu\ell_{\max}\right)$, so the uniform deviation Δ_n has bounded differences with $c_1 = \ldots = c_n = T/n$. Thus, by McDiarmid's inequality [13],

$$\mathcal{L}_{\mathbb{P}_X}(\widetilde{q}_n,\lambda,\mu) \leq \mathcal{L}^*_{\mathbb{P}_X}(\widetilde{q}^*,\lambda,\mu) + 2\,\mathbb{E}\Delta_n + T\sqrt{\frac{2\log(1/\delta)}{n}}, \quad (47)$$

with probability at least $1-\delta$. It remains to upper-bound the expectation $\mathbb{E}\Delta_n$.

Using the identity $\mathbb{E}W = \int_0^T \mathbb{P}(W>t)\mathrm{d}t$ for any random variable W taking values in $[0,T]$ a.s., we can write

$$\mathbb{E}\Delta_n = \mathbb{E}\left[\sup_{g\in\mathcal{G}} \left| \int_0^T \left(\widehat{P}_n\left(g>t\right) - \mathbb{P}_X(g>t)\right)\mathrm{d}t\right|\right] \quad (48)$$

$$\leq T\,\mathbb{E}\left[\sup_{t>0}\sup_{g\in\mathcal{G}} \left|\widehat{P}_n\left(g>t\right) - \mathbb{P}_X(g>t)\right|\right] \quad (49)$$

$$= T\,\mathbb{E}\left[\sup_{A\in\widetilde{\mathcal{A}}_N} \left|\widehat{P}_n(A) - \mathbb{P}_X(A)\right|\right], \quad (50)$$

where $\widetilde{\mathcal{A}}_N$ denotes the collection of all indicator functions of the form $\mathbf{1}_{\{g>t\}}$ for all $g\in\mathcal{G}$ and all $t>0$. A standard estimate from empirical process theory [18] then gives

$$\mathbb{E}\Delta_n \leq CT\sqrt{\frac{\mathrm{V}(\widetilde{\mathcal{A}}_N)}{n}} \quad (51)$$

for some universal constant $C>0$, where $\mathrm{V}(\cdot)$ denotes the VC dimension. Thus, we need to bound $\mathrm{V}(\widetilde{\mathcal{A}}_n)$.

To that end, consider an arbitrary $g = g_{\widetilde{q}} \in \mathcal{G}$ for some $\widetilde{q}\in\widetilde{\mathcal{Q}}(N,L)$. For any $x\in\mathsf{X}$, $g(x)>t$ if and only if

$$d_\mu\left(x,\widetilde{\psi}(s)\right) > t - \lambda\,\mathrm{len}(s), \quad \forall s\in\mathsf{S}. \quad (52)$$

Since $|\mathsf{S}|\leq N$, there are two possibilities to consider for the event $E = \{g_{\widetilde{q}}>t\}$:

1) If $t-\lambda\,\mathrm{len}(s)\geq 0$ for at least one $s\in\mathsf{S}$, then E is an intersection of the complements of at most N sets from the collection $\mathcal{A}$ of all sets of the form (22).
2) If $t<\lambda\,\mathrm{len}(s)$ for all $s\in\mathsf{S}$, then $E=\mathsf{X}$.

This means that $\widetilde{\mathcal{A}}_N \subseteq \bar{\mathcal{A}}_N \cup \{\mathsf{X}\}$, where $\bar{\mathcal{A}}_N$ denotes the collection of all intersections of N complements of sets from $\mathcal{A}$. Then, using Theorem 1.1 in [19], we have

$$\mathrm{V}(\widetilde{\mathcal{A}}_N) \leq 3VN\log(4N), \quad (53)$$

where V is the VC dimension of the class $\mathcal{A}$. Plugging the estimate (53) into (51) and using the resulting bound in (47), we conclude that (23) holds with probability at least $1-\delta$, and the theorem is proved.

REFERENCES

[1] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[2] R. M. Gray, *Source Coding Theory*. Boston: Kluwer, 1990.

[3] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston: Kluwer, 1992.

[4] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Communication, Control and Computing*, 1999, pp. 368–377.

[5] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *J. Machine Learning Res.*, vol. 6, pp. 1705–1749, 2005.

[6] P. Harremoës and N. Tishby, "The Information Bottleneck revisited or how to choose a good distortion measure," in *Proc. IEEE Int. Symp. Inform. Theory*, 2007, pp. 566–570.

[7] S. Lazebnik and M. Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," *IEEE Trans. Pattern Anal. Machine Intelligence*, vol. 31, no. 7, pp. 1294–1309, July 2009.

[8] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "On surrogate loss functions and f-divergences," *Ann. Statist.*, vol. 37, no. 2, pp. 876–904, 2009.

[9] T. Linder, "Learning-theoretic methods in vector quantization," in *Principles of Nonparametric Learning*, L. Györfi, Ed. New York: Springer-Verlag, 2001.

[10] K. R. Parthasarathy, *Probability Measures on Metric Spaces*. New York: Academic Press, 1967.

[11] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Entropy-constrained vector quantization," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 37, no. 1, pp. 31–42, January 1989.

[12] H. S. Witsenhausen, "Indirect rate distortion problems," *IEEE Trans. Inform. Theory*, vol. 26, no. 5, pp. 518–521, 1980.

[13] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.

[14] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Stud. Sci. Math. Hung.*, vol. 2, pp. 299–318, 1967.

[15] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Inform. Theory*, vol. 52, no. 10, pp. 4394–4412, October 2006.

[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.

[17] M. Raginsky, "Joint universal lossy coding and identification of stationary mixing sources with general alphabets," *IEEE Trans. Inform. Theory*, vol. 55, no. 5, pp. 1945–1960, May 2009.

[18] R. M. Dudley, "Central limit theorems for empirical measures," *Ann. Probab.*, vol. 6, pp. 898–929, 1978.

[19] A. van der Vaart and J. A. Wellner, "A note on bounds for VC dimensions," in *High Dimensional Probability V: The Luminy Volume*, C. Houdré, V. Koltchinskii, D. M. Mason, and M. Peligrad, Eds. Beachwood, OH: Institute of Mathematical Statistics, 2009, pp. 103–107.