Quantized Random Projections and Non-Linear Estimation of Cosine Similarity

Authored by:

Ping Li
Martin Slawski
Michael Mitzenmacher

Abstract

Random projections constitute a simple, yet effective technique for dimensionality reduction with applications in learning and search problems. In the present paper, we consider the problem of estimating cosine similarities when the projected data undergo scalar quantization to b bits. We here argue that the maximum likelihood estimator (MLE) is a principled approach to deal with the non-linearity resulting from quantization, and subsequently study its computational and statistical properties. A specific focus is on the on the trade-off between bit depth and the number of projections given a fixed budget of bits for storage or transmission. Along the way, we also touch upon the existence of a qualitative counterpart to the Johnson-Lindenstrauss lemma in the presence of quantization.

1 Paper Body

The method of random projections (RPs) is an important approach to linear dimensionality reduction [23]. RPs have established themselves as an alternative to principal components analysis which is computationally more demanding. Instead of determining an optimal low-dimensional subspace via a singular value decomposition, the data are projected on a subspace spanned by a set of directions picked at random (e.g. by sampling from the Gaussian distribution). Despite its simplicity, this approach comes with a theoretical guarantee: as asserted by the celebrated Johnson-Lindenstrauss (J-L) lemma [6, 12], k = O(log n/?2) random directions are enough to preserve the squared distances between all pairs from a data set of size n up to a relative error of ?, irrespective of the dimension d the data set resides in originally. Inner products are preserved similarly. As a consequence, procedures only requiring distances or inner products can be approximated in the lower-dimensional space, thereby achieving substantial reductions in terms of computation and storage, or mitigating the curse of dimensionality. The idea of RPs has thus been employed

in linear learning [7, 19], fast matrix factorization [24], similarity search [1, 9], clustering [2, 5], statistical testing [18, 22], etc. The idea of data compression by RPs has been extended to the case where the projected data are additionally quantized to b bits so as to achieve further reductions in data storage and transmission. The extreme case of b = 1 is well-studied in the context of locality sensitive hashing [4]. More recently, b-bit quantized random projections for b ? 1 have been considered from different perspectives. The paper [17] studies Hamming distance-based estimation of cosine similarity and linear classification when using a coding scheme that maps a real value to a binary vector of length 2b . It is demonstrated that for similarity estimation, taking b ¿ 1 may yield improvements if the target similarity is high. The paper [10] is dedicated to J-L-type results for quantized RPs, considerably improving over an earlier result of the same flavor in [15]. The work [15] also discusses the trade-off between the number of projections k and number of bits b per projection under a given budget of bits as it also appears in the literature on quantized compressed sensing [11, 14]. In the present paper, all of these aspects and some more are studied for an approach that can be substantially more accurate for small b (specifically, we focus on 1 ? b ? 6) than those in [10, 17, 15]. In [10, 15] the non-linearity of quantization is ignored by treating the quantized data as if they had been observed directly. Such ?linear? approach benefits from its simplicity, but it is geared towards fine quantization, whereas for small b the bias resulting from quantization dominates. By contrast, the approach proposed herein makes full use of the knowledge about the quantizer. As in [17] we suppose that the original data set is contained in the unit sphere of Rd , or at least that the Euclidean 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

norms of the data points are given. In this case, approximating distances boils down to estimating inner products (or cosine similarity) which can be done by maximum likelihood (ML) estimation based on the quantized data. Several questions of interest can be addressed by considering the Fisher information of the maximum likelihood estimator (MLE). With regard to the aforementioned trade-off between k and b, it turns out that the choice b = 1 is optimal (in the sense of yielding maximum Fisher information) as long as the underlying similarity is smaller than 0.2; as the latter increases, the more effective it becomes to increase b. By considering the rate of growth of the Fisher information near the maximum similarity of one, we discover a gap between the finite bit and infinite bit case with rates of ?((1 ? ??)?3/2) and ?((1 ? ??)?2), respectively, where ?? denotes the target similarity. As an implication, an exact equivalent of the J-L lemma does not exist in the finite bit case. The MLE under study does not have a closed form solution. We show that it is possible to approximate the MLE by a non-iterative scheme only requiring pre-computed look-up tables. Derivation of this scheme lets us draw connections to alternatives like the Hamming distance-based estimator in [17]. We present experimental results concerning applications of the proposed approach in nearest neighbor search and linear classification. In nearest neighbor search, we focus on the high similarity regime and confirm theoretical insights into the trade-off between k and

b. For linear classification, we observe empirically that intermediate values of b can yield better trade-offs than single-bit quantization. Notation. We let [d] = {1, . . . , d}. I(P) denotes the indicator function of expression P . For a function f (?), we use f?(?) and f?(?) for its first resp. second derivative. P? and E? denote probability/expectation w.r.t. a zero mean, unit variance bivariate normal distribution with correlation ?. Supplement: Proofs and additional experimental results can be found in the supplement.

2

Quantized random projections, properties of the MLE, and implications

We start by formally introducing the setup, the problem and the approach that is taken before discussing properties of the MLE in this specific case, along with important implications. Setup. Let X = {x1 , . . . , xn } ? Sd?1 , where Sd?1 := {x ? Rd : kxk2 = 1} denotes the unit sphere in Rd , be a set of data points. We think of d being large. As discussed below, the requirement of having all data points normalized to unit norm is not necessary, but it simplifies our exposition considerably. Let x, x0 be a generic pair of elements from X and let ?? = hx, x0 i denote their inner product. Alternatively, we may refer to ?? as (cosine) similarity or correlation. Again for simplicity, we assume that 0 ? ?? ¡ 1; the case of negative ?? is a trivial extension because of symmetry. We aim at reducing the dimensionality of the given data set by means of a random projection, which is realized by sampling a random matrix A of dimension k by d whose entries are i.i.d. N (0, 1) (i.e., zero-mean Gaussian with unit variance). Applying A to X yields Z = {zi }ni=1 ? Rk with zi = Axi , i ? [n]. Subsequently, the projected data points {zi }ni=1 are subject to scalar quantization. A b-bit scalar quantizer is parameterized by 1) thresholds t = (t1 , . . . , tK?1) with 0 = t0 ¡ t1 ¡ . . . ¡ tK?1 ¡ tK = +? inducing a partitioning of the positive real line into K = 2b?1 intervals {[tr?1 , tr), r ? [K]} and 2) a codebook M = {?1 , . . . , ?K } with code ?r representing interval [tr?1 , tr), r ? [K]. Given t and M, the scalar quantizer (or quantization map) is defined by Q : R ? M? := ?M ? M,

z 7? Q(z) = sign(z)
K X
?r I(—z— ? [tr?1 , tr))
(1)
r=1

The projected, b-bit quantized data result as Q = {qi }ni=1 ? (M?)k , qi = (Q(zij))kj=1 , i ? [n]. Problem statement. Let z, z 0 and q, q 0 denote the pairs corresponding to x, x0 in Z respectively Q. The goal is to estimate ?? = hx, x0 i from q, q 0 which automatically yields an estimate of kx ? x0 k22 = 2(1 ? ??). If z, z 0 were given, it would be standard to use k1 hz, z 0 i as an unbiased estimator of ?? . This "linear" approach is commonly adopted when the data undergo uniform quantization with saturation level T (i.e., tr = T ? r/(K ? 1), ?r = (tr ? tr?1)/2, r ? [K ? 1], ?K = T), based on the rationale that as b ? ?, k1 hq, q 0 i ? k1 hz, z 0 i which in turn is sharply concentrated around its expectation ?? . There are two major concerns about this approach. First, for finite b the estimator k1 hq, q 0 i has a bias resulting from the non-

linearity of Q that does not vanish as k ? ?. For small b, the effect of this bias is particularly pronounced. Lloyd-Max quantization (see Proposition 1 below) in place of 2

1.4
p6
p5
p2
0
p1 = P? (Z ? (0, t1], Z ? (0, t1]) p5
p4
p1
p2
p2
p1
p4
p5
p2 = P? (Z ? (0, t1], Z 0 ? (t1 , ?)) p3 = P? (Z ? (t1 , ?), Z 0 ? (t1 , ?))
p3
p2
p5
b=3
1.2
p3
p4 = P?? (Z ? (0, t1], Z 0 ? (0, t1]) p5 = P?? (Z ? (0, t1], Z 0 ? (t1 , ?))
p6
p6 = P?? (Z ? (t1 , ?), Z 0 ? (t1 , ?))
1
empirical MSE I?1(?)
0.8 0.6 0.4 0.2 0 0
0.2
0.4
0.6
0.8
1
?

Figure 1: (L, M): Partitioning into cells for b = 2 and cell probabilities. (R): Empirical MSE k(b ?MLE ? ??)2 for b = 3 (averaged over 104 i.i.d. data sets with k = 100) compared to the inverse information. The disagreement for ? ? 0.2 results from positive truncation of the MLE at zero. uniform quantization provides some remedy, but the issue of non-vanishing bias remains. Second, even for infinite b, the approach is statistically not efficient. In order to see this, note that

1 ?? i.i.d. {(zj , zj0)}kj=1 ? (Z, Z 0), where (Z, Z 0) ? N2 0, . (2) ?? 1 It is shown in [16] that the MLE of ?? under the above bivariate normal model has a variance of (1 ? ?2?)2 /{k (1 + ?2?)}, while Var(hz, z 0 i /k) = (1 + ?2?)/k which is a substantial difference for large ?? . The higher variance

results from not using the information that the components of z and z 0 have unit variance [16]. In conclusion, the linear approach as outlined above suffers from noticeable bias/and or high variance if the similarity ?? is high, and it thus makes sense to study alternatives. Maximum likelihood estimation of ?? . We here propose the MLE in place of the linear approach. The advantage of the MLE is that it can have substantially better statistical performance as the quantization map is explicitly taken into account. The MLE is based on bivariate normality according to (2). The effect of quantization is identical to that of what is known as interval censoring in statistics, i.e., in place of observing a specific value, one only observes that the datum is contained in an interval. The concept is easiest to understand in the case of one-bit quantization. For any j ? [k], each of the four possible outcomes of (qj , qj0) corresponds to one of the four orthants of R2 . By symmetry, the probability of (qj , qj0) falling into the positive or into the negative orthant are identical; both correspond to a ?collision?, i.e., to the event {qj = qj0 }. Likewise, the probability of (qj , qj0) falling into one of the remaining two orthants are identical, corresponding to a disagreement {qj 6= qj0 }. Accordingly, the likelihood function in ? is given by k Y

0

0

{?(?)I(qj =qj) (1 ? ?(?))I(qj 6=qj) },

?(?) := P? (sign(Z) = sign(Z 0)),

j=1

where ?(?) denotes the probability of a collision after quantization for (Z, Z 0) as in (2) with ?? replaced by ?. It is straightforward to show that the MLE is given by ?bMLE = cos(?(1 ? ? b)), where Pk ? is the circle constant and ? b = k ?1 j=1 I(qj = qj0) is the empirical counterpart to ?(?). We note that the expression for ?bMLE follows the same rationale as used for the simhash in [4]. With these preparations, it is not hard to see how the MLE generalizes to cases with more than one bit. For b = 2, there is a single non-trivial threshold t1 that yields a partitioning of the real axis into four bins and accordingly a component (qj , qj0) of a quantized pair can fall into 16 possible cells (rectangles), cf. Figure 1. By orthant symmetry and symmetries within each orthant, one ends up with six distinct probabilities p1 , . . . , p6 for (qj , qj0) falling into one of those cells depending on ?. Weighting those probabilities according to the number of their occurrences in the left part of Figure 1, we end up with probabilities ?1 = ?1 (?), . . . , ?6 = ?6 (?) that sum up to one. The corresponding relative cell frequencies ? b1 , . . . , ? b6 resulting from (qj , qj0)kj=1 form a sufficient statistic for ?. For 2b general b, we have 2 cells and L = K(K + 1) (recall that K = 2b?1) distinct probabilities, so that L = 20, 72, 272, 1056 for b = 3, . . . , 6. This yields the following compact expressions for the 3

1.8

2

2.5

b=4

b=2

b=6
1.8
1.6 1.6
1
0.6 0.4 0
b ? Ib?1 (?)/I1?1 (?)
1.2
1.2
0.8
2
1.4
b ? Ib?1 (?)/I1?1 (?)
b ? Ib?1 (?)/I1?1 (?)
1.4
1.5
1
0.8
Lloyd-Max T0.9 T0.95 T0.99 0.2
Lloyd-Max T0.9 T0.95 T0.99
0.6 0.4 0.2
0.4
0.6
0.8
1
0 0
?
0.2
1
0.5
0.4
0.6
0.8
1
Lloyd-Max T0.9 T0.95 T0.99
0 0
0.2
0.4
0.6
?
0.8
1
?

Figure 2: b ? Ib?1 (?)/I1?1 (?) vs. ? for different choices of t: Lloyd-Max and uniform quantization with saturation levels T0.9 , T0.95 , T0.99 , cf. ?4.1 for a definition. The latter are better suited for high similarity. The differences become smaller as b increases. Note that for b = 6, ? ¿ 0.7 is required for

either quantization scheme to achieve a better trade-off than the one-bit MLE. negative log-likelihood l(?) and the Fisher information I(?) = E? [?l(?)] (up to a factor of k) l(?) =

L X
? b‘ log(¿ (?)),
‘=1
I(?) =
L X (?? ‘ (?))2 ‘=1
¿ (?)
.
(3)

The information I(?) is of particular interest. By classical statistical theory [21], {E[b ?MLE] ? ?? }2 = 2 ?1 2 ?1 2 O(1/k), Var(b ?MLE) = I (?)/k, E[(b ?MLE ? ??)] = I (?)/k + O(1/k) as k ? ?. While this is an asymptotic result, it agrees to a good extent with what one observes for finite, but not too small samples, cf. Figure 1. We therefore treat the inverse information as a proxy for the accuracy of ?bMLE in subsequent analysis. Remark. We here briefly address the case of known, but possibly non-unit norms, i.e., kxk2 = ?x , kx0 k2 = ?x0 . This can be handled by re-scaling the thresholds of the quantizer (1) by ?x resp. ?x0 , estimating ?? based on q, q 0 as in the unit norm case, and subsequently re-scaling the estimate by ?x ?x0 to obtain an estimate of hx, x0 i. The assumption that the norms are known is not hard to satisfy in practice as they can be computed by one linear scan during data collection. With a limited bit budget, the norms additionally need to be quantized. It is unclear how to accurately estimate them from quantized data (for b = 1, it is definitely impossible). Choice of the quantizer. Equipped with the Fisher information (3), one of the questions that can be addressed is quantizer design. Note that as opposed to the linear approach, the specific choice of the {?r }K r=1 in (1) is not important as ML estimation only depends on cell frequencies but not on the values associated with the intervals {(tr?1 , tr]}K r=1 . The thresholds t, however, turn out to have a considerable impact, at least for small b. An optimal set of thresholds can be determined by minimizing the inverse information I ?1 (?; t) w.r.t. t for fixed ?. As the underlying similarity is not known, this may not seem practical. On the other hand, prior knowledge about the range of ? may be available, or the closed form one-bit estimator can be used as pilot estimator. For ? = 0, the optimal set of thresholds coincide with those of Lloyd-Max quantization [20]. Proposition 1. Let g ? N (0, 1) and consider Lloyd-Max quantization given by K 2 ? ?1 (t? , {??r }K (0; t). r=1) = argmin E[{g ? Q(g; t, {?r }r=1)}]. We also have t = argmin I t
t,{?r }K r=1

The Lloyd-Max problem can be solved numerically by means of an alternating scheme which can be shown to converge to a global optimum [13]. For ? ¿ 0, an optimal set of thresholds can be determined by general procedures for nonlinear optimization. Evaluation of I ?1 (?; t) requires computation of the probabilities {¿ (?; t)}L ? ‘ (?; t)}L ‘=1 and their derivatives {? ‘=1 . The latter are available in closed form (cf. supplement), while for the former

specialized numerical integration procedures [8] can be used. In order to avoid multi-dimensional optimization, it makes sense to confine oneself to thresholds of the form tr = T ? r/(K ? 1), r ? [K ? 1], so that only T needs to be optimized. Even though the Lloyd-Max scheme performs reasonably also for large values of ?, the one-parameter scheme may still yield significant improvements in that case, cf. Figure 2. Once b ? 5, the differences between the two schemes become marginal. Trade-off between k and b. Suppose we are given a fixed budget of bits B = k ? b for transmission or storage, and we are in free choosing b. The optimal choice of b can be determined by comparing 4

7
0.25
1 2 3 4 5 6
b b b b b b
? ? 0.9 0.2
b ? Ib?1 (?)
b ? Ib?1 (?)
5
= = = = = =
4 3
0.15
= = = = = =
6
1 2 3 4 5 6
5
optimal b
b b b b b b
all ? 6
0.1
4
3
2 0.05
2
1 0 0
0.2
0.4
0.6
0.8
1
0 0.9
0.92
0.94
?
0.96
?
0.98
1

1 0
0.2
0.4
?
0.6
0.8
1
b ? Ib?1 (?)

Figure 3: Trade-off between k and b. (L): vs. ? for 1 ? b ? 6 with t chosen by Lloyd-Max. (M): Zoom into the range 0.9 ? ? ? 1. (R): choice of b minimizing b ? Ib?1 (?) vs. ?. the inverse Fisher information Ib?1 (?) for changing b with t chosen according to either of the two schemes above. Since the mean squared error of ?bMLE decays with 1/k for any b, for b0 with b0 ¿ b to be more efficient than b at the bit scale it, is required that Ib0 (?)/Ib (?) ¿ b0 /b as with the smaller choice b one would be allowed to increase k by a factor of b0 /b. Again, this comparison is dependent on a specific ?. From Figure 3, however, one can draw general conclusions: for ? ¡ 0.2, it does not pay off to increase b beyond one; as ? increases, higher values of b achieve a better trade-off with even b = 6 being the optimal choice for ? ¿ 0.98. The intuition is that two points of high similarity agree on their first significant bit for most coordinates, in which case increasing the number of bits becomes beneficial. This finding is particularly relevant to (near-)duplicate detection/nearest neighbor search where high similarities prevail, an application investigated in ?4. Rate of growth of the Fisher information near ? = 1. Interestingly, we do not observe a ?saturation? even for b = 6 in the sense that for ? close enough to 1, one can still achieve an improvement at the bit scale compared to 1 ? b ? 5. This raises the question about the rate of growth of the Fisher information near one relative to the full precision case (b ? ?). As shown in [16] I? (?) = (1 + ?2)/(1 ? ?2)2 = ?((1 ? ?)?2) as ? ? 1. As stated below, in the finite bit case, the exponent is only 3/2 for all b. This is a noticeable gap. Theorem 1. For 1 ? b ¡ ?, we have I(?) = ?((1 ? ?)?3/2) as ? ? 1. The theorem has an interesting implication with regard to the existence of a Johnson-Lindenstrauss (J-L)-type result for quantized random projections. In a nutshell, the J-L lemma states that as long as k = ?(log n/?2), with high probability we have that (1 ? ?)kxi ? xj k22 ? kzi ? zj k22 /k ? (1 + ?)kxi ? xj k22 for all pairs (i, j), i.e., the distances of the data in X are preserved in Z up to a relative error of ?. In our setting, one would hope for an equivalent of the form 2 (1 ? ?)2(1 ? ?ij) ? 2(1 ? ?bij MLE) ? (1 + ?)2(1 ? ?ij) ?(i, j) as long as k = ?(log n/?), (4)

where ?ij = hxi , xj i, i, j ? [n], and ?bij MLE denotes the MLE for ?ij given quantized RPs. The standard proof of the J-L lemma [6] combines norm preservation for each individual pair of the form P((1 ? ?)kxi ? xj k22 ? kzi ? zj k22 /k ? (1 + ?)kxi ? xj k22) ? 2 exp(?k?(?2)) with a union bound. Such a concentration result does not appear to be attainable for ?bMLE ? ?? , not even asymptotically as k ? ? in which case ?bMLE ? ?? is asymptotically normal with mean zero and variance I ?1 (??)/k. This yields an asymptotic tail bound of the form P(—b ?MLE ? ?? — ¿ ?) ? 2 exp(?? 2 k/{2I ?1 (??)}). (5) For a

result of the form (4), which is about relative distance preservation, one would need to choose ? proportional to ?(1 ? ??). In virtue of Theorem 1, I ?1 (??) = ?((1 ? ??)3/2) as ?? ? 1 so that with ? chosen in that way the exponent in (5) would vanish as ?? ? 1. By constrast, the required rate of decay of I ?1 (??) is achieved in the full precision case. Given the asymptotic optimality of the MLE according to the Cramer-Rao lower bound suggests that a qualitative counterpart to the J-L lemma (4) is out of reach. Weaker versions in which the required lower bound on k would depend inversely on the minimum distance of points in X are still possible. Similarly, a weaker result of the form 2 2(1 ? ?ij) ? ? ? 2(1 ? ?bij MLE) ? 2(1 ? ?ij) + ? ?(i, j) as long as k = ?(log n/?), is known to hold already in the one-bit case and follows immediately from the closed form expression of the MLE, Hoeffdings?s inequality, and the union bound; cf. e.g. [10].

5

3

A general class of estimators and approximate MLE computation

A natural concern about the MLE relative to the linear approach is that it requires optimization via an iterative scheme. The optimization problem is smooth, one-dimensional and over the unit interval, hence not challenging for modern solvers. However, in applications it is typically required to compute the MLE many times, hence avoiding an iterative scheme for optimization is worthwhile. In this section, we introduce an approximation to the MLE that only requires at most two table look-ups. PL A general class of estimators. Let ?(?) = (?1 (?), . . . , ?L (?))¿ , ‘=1 ¿ (?) = 1, be the normalized cell frequencies depending on ? as defined in ?2, let further w ? RL be a fixed vector of weights, and consider the map ? 7? ?(?; w) := h?(?), wi. If h?(?), ? wi ¿ 0 uniformly in ? (such w always exist), ?(?; w) is increasing and has an inverse ??1 (? ; w). We can then consider the estimator ?bw = ??1 (hb ? , wi ; w),

(6)

where we recall that ? b = (b ?, . . . , ? bL)¿ are the empirical cell frequencies given quantized data q, q 0 . It is easy to see that ?bw is a consistent estimator of ?? : we have ? b ? ?(??) in probability by the law of large numbers, and ??1 (hb ? , wi ; w) ? ??1 (h?(??), wi ; w) = ??1 (?(?? ; w); w) = ?? by two-fold application of the continuous mapping theorem. By choosing w such that w‘ = 1 for ‘ corresponding to cells contained in the positive/negative orthant and w‘ = ?1 otherwise, ?bw becomes the one-bit MLE. By choosing w‘ = 1 for diagonal cells (cf. Figure 1) corresponding to a collision event {qj = qj0 } and w‘ = 0 otherwise, we obtain the Hamming distance-based estimator in [17]. Alternatively, we may choose w such that the asymptotic variance of ?bw is minimized. Theorem 2. For any w s.t. ?(? ? ?)¿ w 6= 0, we have Var(b ?w) = V (w; ??)/k + O(1/k 2) as k ? ?, V (w; ??) = (w¿ ?(??)w)/{?(? ? ?)¿ w}2 , ?(??) := ?(??) ? ?(??)?(??)¿ , ? ?1 and ?(??) := diag((¿ (??))L (??)?(? ? ?). Then: ‘=1). Moreover, let w = ? argminw V (w; ??) = {?(w? + c1), ? 6= 0, c ? R}, and E[(b ?w? ? ??)2] = E[(b ?MLE ? ??)2] + O(1/k 2).

V (w? ; ??) = I ?1 (??),

Theorem 2 yields an expression for the optimal weights w? = ??1 (??)?(? ? ?). This optimal choice is unique up to translation by a multiple of the constant vector 1 and scaling. The estimator ?bw? based on the choice w = w? achieves asymptotically the same statistical performance as the MLE. Approximate computation. The estimator ?bw? is not operational as the optimal choice of the weights depends on the estimand itself. This issue can be dealt with by using a pilot estimator ?b0 like R1 the one-bit MLE, the Hamming distance-based estimator in [17] or ?b0 = ?bw , where w = 0 w(?) d? averages the expression w(?) = ??1 (?)?(?) ? for the optimal weights over ?. Given the pilot estimator, we may then replace w? by w(b ?0) and use ?bw(b?0) as a proxy for ?bw? which achieves the same statistical performance asymptotically. A second issue is that computation of ?bw (6) entails inversion of the function ?(?; w). The inverse may not be defined in general, but for the choices of w that we have in mind, this is not a concern (cf. supplement). Inversion of ?(?; w) can be carried out with tolerance ? by tabulating the function values on a uniform grid of cardinality d1/?e and performing a table lookup for each query. When computing ?bw(b?0) , the weights depends on the data via the pilot estimator. We thus need to tabulate w(?) on a grid, too. Accordingly, a whole set of look-up tables is required for function inversion, one for each set of weights. Given parameters ?, ? ¿ 0, a formal description of our scheme is as follows. 1. Set R = d1/?e, ?r = r/R, r ? [R], and B = d1/?e, ?b = b/B, b ? [B]. 2. Tabulate w(?b), b ? [B], and function values ?(?r ; w(?b)) = hw(?b), ?(?r)i, r ? [R], b ? [B]. Steps 1. and 2. constitute a one-time pre-processing. Given data q, q 0 , we proceed as follows. 3. Obtain ? b and the pilot estimator ?b0 = ??1 (hb ? , wi ; w), with w defined in the previous paragraph. 4. Return ?b = ??1 (hb ? , w(e ?0)i ; w(e ?0)), where ?e0 is the value closest to ?b0 among the {?b }. Step 2. requires about C = d1/?e ? d1/?e ? L computations/storage. From experimental results we find that ? = 10?4 and ? = .02 appear sufficient for practical purposes, which is still manageable even for b = 6 with L = 1056 cells in which case C ? 5 ? 108 . Again, this cost is occurred only once independent of the data. The function inversions in steps 3. and 4. are replaced by table lookups. By organizing computations efficiently, the frequencies ? b can be obtained from one pass over (qj ? qj0), j ? [k]. Equipped with the look-up tables, estimating the similarity of two points requires O(k + L + log(1/?)) flops which is only slightly more than a linear scheme with O(k). 6

1

1

1 0.95

0.95

0.9

0.8

0.7

0.6

synthetic, K = 10 0.5

6

7

8
9 10 log2(bits)
b=1 b=2 b=3 b=4 b=5 b=6 b=? oracle 11
12
0.85 0.8
b=1 b=2 b=3 b=4 b=5 b=6 b=? oracle
0.75 0.7 0.65
farm, K = 10 13
fraction retrieved
fraction retrieved
fraction retrieved
0.9
0.6
6
7
8
9 10 log2(bits)
11
12
0.9
0.85
b=1 b=2 b=3 b=4 b=5 b=6 b=? oracle
0.8
0.75
rcv1, K = 10 13
0.7
6
7
8
9 10 log2(bits)
11
12
13

Figure 4: Average fraction of K = 10 nearest neighbors retrieved vs. total # of bits (log2 scale) for 1 ? b ? 6. b = ? (dashed) represents the MLE based on unquantized data, with k as for b = 6. The oracle curve (dotted) corresponds to b = ? with maximum k (i.e., as for b = 1).

4

Experiments

We here illustrate the approach outlined above in nearest neighbor search and linear classification. The focus is on the trade-off between b and k, in particular in the presence of high similarity. 4.1 Nearest Neighbor Search Finding the most similar data points for a given query is a standard task in information retrieval. Another application is nearest neighbor classification. We here investigate how the performance of our approach is affected by the choice of k,

b and the quantization scheme. Moreover, we compare to two baseline competitors, the Hamming distance-based approach in [17] and the linear approach in which the quantized data are treated like the original unquantized data. For the approach in [17], Pk similarity of the quantized data is measured in terms of their Hamming distance j=1 I(qj 6= qj0). Synthetic data. We generate k i.i.d. samples of Gaussian data, where each sample X = (X0 , X1 , . . . , X96) is generated as X0 ? N (0, 1), Xj = ?j X0 + (1 ? ?2j)1/2 Zj , 1 ? j ? 96, where 2 the {Zj }96 j=1 are i.i.d. N (0, 1) and independent of X0 . We have E[(X0 ? Xj)] = 2(1 ? ?j), where ?j = min{0.8+(j ?1)0.002, 0.99}, 1 ? j ? 96. The thus generated data subsequently undergo b-bit quantization, for 1 ? b ? 6. Regarding the number of samples, we let k ? {26 /b, 27 /b, . . . , 213 /b} which yields bit budgets between 26 and 213 for all b. The goal is to recover the K nearest neighbors of X0 according to the {?j }, i.e., X96 is the nearest neighbor etc. The purpose of this specific setting is to mimic the use of quantized random projections in the situation of a query x0 and data points X = {x1 , . . . , x96 } having cosine similarities {?j }96 j=1 with the query. Real data. We consider the Farm Ads data set (n = 4, 143, d = 54, 877) from the UCI repository and the RCV1 data set (n = 20, 242, d = 47, 236) from the LIBSVM webpage [3]. For both data sets, each instance is normalized to unit norm. As queries we select all data points whose first neighbor has (cosine) similarity less than 0.999, whose tenth neighbor has similarity at least 0.8 and whose hundredth neighbor has similarity less than 0.5. These restrictions allow for a more clear presentation of our results. Prior to nearest neighbor search, b-bit quantized random projections are applied to the data, where the ranges for b and for the number of projections k is as for the synthetic data. Quantization. Four different quantization schemes are considered: Lloyd-Max quantization and thresholds tr = T? ? r/(K ? 1), r ? [K ? 1], where T? is chosen to minimize I ?1 (?); we consider ? ? {0.9, 0.95, 0.99}. For the linear approach, we choose ?r = E[g—g ? (tr?1 , tr)], r ? [K], where g ? N (0, 1). For our approach and that in [17] the specific choice of the {?r } is not important. Evaluation. We perform 100 respectively 20 independent replications for synthetic respectively real data. We then inspect the top K neighbors for K ? {3, 5, 10} returned by the methods under consideration, and for each K we report the average fraction of true K neighbors that have been retrieved over 100 respectively 20 replications, where for the real data, we also average over the chosen queries (366 for farm and 160 for RCV1). The results of our experiments point to several conclusions that can be summarized as follows. One-bit quantization is consistently outperformed by higher-bit quantization. The optimal choice of b depends on the underlying similarities, and interacts with the choice of t. It is an encouraging result that the performance based on full precision data (with k as for b = 6) can essentially be matched 7

0.95
1
0.9
0.9
1

1
0.9
0.75
fraction retrieved
fraction retrieved
fraction retrieved
0.8
0.8
0.7
farm, b = 4, K = 10 0.6
fraction retrieved
0.95 0.85
0.9
0.85
0.8
0.7
0.6
0.7 0.65 0.6
MLE Hamming Linear
farm, b = 2, K = 10
0.4 6
7
8
9 10 log2(bits)
11
12
13
6
7
8
9 10 log2(bits)
11
12
rcv1, b = 4, K = 10
0.8
MLE Hamming Linear
0.5
MLE Hamming Linear
rcv1, b = 2, K = 10
13
0.75
0.4 6
7
8
9 10 log2(bits)
11

12
MLE Hamming Linear
0.5
13
6
7
8
9 10 log2(bits)
11
12
13

Figure 5: Average fraction of K = 10 nearest neighbors retrieved vs. total # of bits (log2 scale) of our approach (MLE) relative to that based on the Hamming distance and the linear approach for b = 2, 4. when quantized data is used. For b = 2, the performance of the MLE is only marginally better than the approach based on the Hamming distance. The superiority of the former becomes apparent once b ? 4 which is expected since for increasing b the Hamming distance is statistically inefficient as it only uses the information whether a pair of quantized data agrees/disagrees. Some of these findings are reflected in Figures 4 and 5. We refer to the supplement for additional figures. 4.2 Linear Classification We here outline an application to linear classification given features generated by (quantized) random b= projections. We aim at reconstructing the original Gram matrix G = (hxi , x0i i)1?i,i0 ?n from G (b gii0), where for i 6= i0 , gbii0 = ?bMLE (qi , qi0) equals the MLE of hxi , x0i i given a quantized data pair b is subsequently fed into LIBSVM. qi , qi0 , and gbii0 = 1 else (assuming normalized data). The matrix G For testing, the inner products between test and training pairs are approximated accordingly. Setup. We work with the farm data set using the first 3,000 samples for training, and the Arcene data set from the UCI repository with 100 training and 100 test samples in dimension d = 104 . The choice of k and b is as in ?4.1; for arcene, the total bit budget is lowered by a factor of 2. We perform 20 independent replications for each combination of k and b. For SVM classification, we consider logarithmically spaced grids between 10?3 and 103 for the parameter C (cf. LIBSVM manual).

0.85
farm b=1 b=2 b=3 b=4 b=5 b=6 b=? oracle
0.8
0.75
0.7 8
9
10 11 log2(bits)
12
0.85
arcene
0.8
b=1 b=2 b=3 b=4 b=5 b=6 b=? oracle

0.75
0.7
13
7
8
9 10 log2(bits)
11
12
accuracy on test set
0.85
accuracy on test set
accuracy on test set
0.9
0.8
0.75
b=1 b=2 b=3 b=4 b=5 b=6 b=? oracle
0.7
0.65
arcene, total #bits = 210 0
0.5
1 1.5 log10(C parameter)
2
2.5

Figure 6: (L, M): accuracy vs. bits, optimized over the SVM parameter C. (R) accuracy vs. C for a fixed # bits. b = ? indicates the performance based on unquantized data with k as for b = 6. The oracle curve (dotted) corresponds to b = ? with maximum k (i.e., as for b = 1). Figure 6 (L, M) displays the average accuracy on the test data (after optimizing over C) in dependence of the bit budget. For the farm Ads data set, b = 2 achieves the best trade-off, followed by b = 1 and b = 3. For the Arcene data set, b = 3, 4 is optimal. In both cases, it does not pay off to go for b ? 5.

5

Conclusion

In this paper, we bridge the gap between random projections with full precision and random projections quantized to a single bit. While Theorem 1 indicates that an exact counterpart to the J-L lemma is not attainable, other theoretical and empirical results herein point to the usefulness of the intermediate cases which give rise to an interesting trade-off that deserves further study in contexts where random projections can naturally be applied e.g. linear learning, nearest neighbor classification or clustering. The optimal choice of b eventually depends on the application: increasing b puts an emphasis on local rather than global similarity preservation.

8

Acknowledgement The work of Ping Li and Martin Slawski is supported by NSF-Bigdata-1419210 and NSF-III-1360971. The work of Michael Mitzenmacher is supported by NSF CCF-1535795 and NSF CCF-1320231.

2 References

[1] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In Conference on Knowledge discovery and Data mining (KDD), pages 245?250, 2001. [2] C. Boutsidis, A. Zouzias, and P. Drineas. Random Projections for k-means Clustering. In Advances in Neural Information Processing Systems (NIPS), pages 298?306. 2010. [3] C-C. Chang and C-J. Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1?27:27, 2011. http://www.csie.ntu.edu.tw/cjlin/libsvm. [4] M. Charikar. Similarity estimation techniques from rounding algorithms. In Proceedings of the Symposium on Theory of Computing (STOC), pages 380?388, 2002. [5] S. Dasgupta. Learning mixtures of Gaussians. In FOCS, pages 634?644, 1999. [6] S. Dasgupta. An elementary proof of a theorem of Johnson and Lindenstrauss. Random Structures and Algorithms, 22:60?65, 2003. [7] D. Fradkin and D. Madigan. Experiments with random projections for machine learning. In Conference on Knowledge discovery and Data mining (KDD), pages 517?522, 2003. [8] A. Genz. BVN: A function for computing bivariate normal probabilities. http://www.math.wsu.edu/ faculty/genz/homepage. [9] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In Proceedings of the Symposium on Theory of Computing (STOC), pages 604?613, 1998. [10] L. Jacques. A Quantized Johnson-Lindenstrauss Lemma: The Finding of Buffon?s needle. IEEE Transactions on Information Theory, 61:5012?5027, 2015. [11] L. Jacques, K. Degraux, and C. De Vleeschouwer. Quantized iterative hard thresholding: Bridging 1-bit and high-resolution quantized compressed sensing. arXiv:1305.1786, 2013. [12] W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. Contemporary Mathematics, pages 189?206, 1984. [13] J. Kieffer. Uniqueness of locally optimal quantizer for log-concave density and convex error weighting function. IEEE Transactions on Information Theory, 29:42?47, 1983. [14] J. Laska and R. Baraniuk. Regime change: Bit-depth versus measurement-rate in compressive sensing. IEEE Transactions on Signal Processing, 60:3496?3505, 2012. [15] M. Li, S. Rane, and P. Boufounos. Quantized embeddings of scale-invariant image features for mobile augmented reality. In International Workshop on Multimedia Signal Processing (MMSP), pages 1?6, 2012. [16] P. Li, T. Hastie, and K. Church. Improving Random Projections Using Marginal Information. In Annual Conference on Learning Theory (COLT), pages 635?649, 2006. [17] P. Li, M. Mitzenmacher, and A. Shrivastava. Coding for Random Projections. In Proceedings of the International Conference on Machine Learning (ICML), 2014. [18] M. Lopes, L. Jacob, and M. Wainwright. A More Powerful Two-Sample Test in High Dimensions using Random Projection. In Advances in Neural Information Processing Systems 24, pages 1206?1214. 2011. [19] O. Maillard and R. Munos. Compressed least-squares regression. In Advances in Neural Information Processing Systems (NIPS), pages 1213?1221. 2009. [20] J. Max. Quantizing for Minimum Distortion. IRE Transactions on Information Theory, 6:7?12, 1960. [21] L. Shenton and K. Bowman. Higher Moments of a

Maximum-likelihood Estimate. Journal of the Royal Statistical Society, Series B, pages 305?317, 1963. [22] R. Srivastava, P. Li, and D. Ruppert. RAPTT: An exact two-sample test in high dimensions using random projections. Journal of Computational and Graphical Statistics, 25(3):954?970, 2016. [23] S. Vempala. The Random Projection Method. American Mathematical Society, 2005. [24] F. Wang and P. Li. Efficient nonnegative matrix factorization with random projections. In SDM, pages 281?292, Columbus, Ohio, 2010.

9