Approximation Algorithms Workshop June 13-17, 2011, Princeton

Why Do We Want a Good Ratio Anyway? Approximation Stability and Proxy Objectives

Scribe: Avrim Blum *Avrim Blum*

1 Overview

When real-world problems are abstracted as optimization problems, it is often the case that the formal objective used in the optimization problem is serving as a proxy for some other underlying goal. For example, if we have a clustering problem such as clustering proteins by function, we might represent our data (protein sequences) in some natural way as points in a metric space, and then abstract this as a k-median problem where we aim to find k clusters, along with a center point for each cluster, such that the sum of distances of the data points to their cluster centers is minimized. Here, the k-median objective is serving as a proxy. Our hope is that we have represented data in such a way that a good k-median solution will translate to a good solution to the true goal (having clusters that actually match the proteins' functions).

In the field of Approximation Algorithms, we typically ignore this latter aspect and do not aim to model it: instead, we consider the optimization problem as given, and aim to understand the best worst-case approximation ratios achievable. Here, we consider: what happens if we incorporate the connection between the two objectives into the theory? That is, suppose that the *reason* we want a good approximation ratio is that we believe that our instance satisfies the *promise* that a good approximation to the formal objective (e.g., k-median) will imply a near-optimal solution to our underlying goal (e.g., having our clustering be a good match to an unknown target clustering). Does this allow us to achieve better guarantees, and perhaps get around what on worst-case instances would be intrinsic computational barriers?

What we find is that for certain problems, this condition, which we will call *approximation-stability*, yields a number of structural properties of the instance that can indeed be used algorithmically. For example, continuing with the case of clustering, the best k-median approximation guarantee known for worst-case instances is a factor $3+\epsilon$ [AGK$^+$04] (building on a long line of work initiated by [CGTS99]), and it is known to be NP-hard to beat a factor of $1+1/e$ [JMS02].[1] Therefore, one would think that in order to produce a guarantee on the quality of the solution found, one would need (for the current best algorithms) an assumption that applied to 3-approximations. Further, one would think that a promise such as "on this instance, achieving a 1.1-approximation to the k-median optimal is sufficient to be ϵ-close to the unknown target clustering" would be useless, given the NP-hardness of achieving a 1.1-approximation in general. However, this is not the case. In fact, for both the k-median and k-means problems, it is shown in [BBG09] that one can devise algorithms that will efficiently find clusterings that are $O(\epsilon)$-close to the unknown target clustering under this promise; similar guarantees for the min-sum objective are given in [BB09]. Further, for k-median and k-means objectives, it is shown in [ABS10] that if target clusters are larger than size

[1] Throughout this text, k should *not* be viewed as a constant (an $O(n^k)$-time exact algorithm is trivial for k-median, for instance) and instead as some function of n such as $n^{0.1}$.

ϵn (for this to be interesting, one should view ϵ as sub-constant) then under this condition one can in fact obtain a PTAS (which in turn, by the promise, would yield a solution that is ϵ-close to the target). These results suggest that in cases where the objective function is a proxy for some other goal, such as matching a target clustering, modeling the connection between the two goals may be able to provide a theoretically-analyzable avenue around worst-case hardness results.

Let us now define approximation-stability more generally.[2]

Definition 1 (Approximation Stability)**.** *An instance having some unknown correct solution $\mathcal{C}^*$ satisfies* (c, ϵ)-approximation stability *with respect to objective function Φ and distance measure dist, if all solutions $\mathcal{C}$ with $\Phi(\mathcal{C}) \leq c \cdot \Phi(\mathbf{OPT})$ satisfy $dist(\mathcal{C}, \mathcal{C}^*) \leq \epsilon$. Here, $\mathbf{OPT}$ is the optimal solution under objective Φ, which need not necessarily be the same as $\mathcal{C}^*$.*

In the case of clustering, a "solution" is a clustering and the natural notion of distance is the fraction of points would have to be reassigned in one clustering in order to make it equal to the other.

Approximation stability is in essence a promise condition: a promise that any sufficiently good approximation to the given objective Φ will be close to a solution one is looking for. It can be viewed as a formal motivation for aiming to approximate the measurable objective Φ (such as k-means) when the ultimate goal is to match some desired solution $\mathcal{C}^*$ well. To put it another way, if the condition does *not* hold, then one would want to investigate what sort of properties one should aim for in a solution *beyond* having a near-optimal objective value.

Clearly if an instance satisfies (c, ϵ)-approximation stability for a value c that is greater than or equal to the best approximation guarantee known for the problem, then we can immediately find a solution that is ϵ-close to the target (namely, just run the approximation algorithm). So, the interesting case is when c is less than the best known guarantee, or even when it is below a hardness threshold.

As mentioned above, for clustering under the k-median, k-means, or min-sum objectives, we have algorithms [BBG09, BB09] that can use (c, ϵ)-approximation stability to find solutions that are $O(\epsilon)$-close to the desired solution, for any constant $c > 1$. An interesting open question is to consider other clustering-based objective functions. For example, for the *sparsest cut* problem, the best approximation guarantee known is a factor $O(\sqrt{\log n})$ [ARV04]. Suppose an instance satisfies (c, ϵ)-approximation stability for $c = 100$: is this sufficient to find either an $O(1)$ approximation to the objective, or (as above) a solution that differs in only an $O(\epsilon)$ fraction of points compared to a target partitioning?

One can also consider other kinds of problems. For example [ABB$^+$10, BB11] consider the problem of finding approximate Nash equilibria. In this context, approximation-stability means that all approximate equilibria should be close (as probability distributions) to some unknown true equilibrium, motivated by settings where the goal is to predict behavior of players. Another natural setting in which to examine approximation-stability would be the problem of phylogenetic tree reconstruction. Here, one typically defines a Steiner-tree-like objective function, and yet the true goal is to match an unknown correct evolutionary tree.

[2] In the conference version of [BBG09] this was called the "(c, ϵ) property for objective Φ". This was changed to the more descriptive term "approximation stability" in the long version and in subsequent papers.

2 More information

The following are papers that analyze approximation-stability for a number of different problems, as well as papers that experimentally examine the performance of algorithms designed for approximation-stability in real applications.

[BBG09] This work introduced the notion of approximation-stability, and gave results for clustering instances stable to k-median, k-means, and min-sum objectives. For k-median and k-means, it shows that $(1+\alpha, \epsilon)$-approximation stability is sufficient to efficiently produce a clustering of distance $O(\epsilon/\alpha)$ from the target clustering; for the min-sum objective, the same conclusion is reached under the additional assumption that all target clusters are large compared to $\epsilon n/\alpha$. This work also showed that when all target clusters are large compared to $\epsilon n/\alpha$, for the k-median objective one can produce a clustering of distance $\leq \epsilon$ from the target (i.e., as good as if one had a $1+\alpha$ approximation algorithm).

[BB09] This work gives improved results for the min-sum objective, removing the size restriction on clusters needed by [BBG09], as well as presenting additional improvements. This work also presents results for the *correlation-clustering* objective.

[VBR$^+$10, VBR$^+$11] These papers consider the important problem of clustering protein sequences by function, giving both theoretical and experimental results. First, due to computational constraints in this setting, these papers consider a framework in which (approximate) pairwise distance information between data points can only be obtained via conducting a small number of "one versus all" queries. [VBR$^+$10] then develops a fast algorithm that operates under these constraints and, building on the analysis of [BBG09], proves that it achieves strong guarantees under k-median approximation-stability. [VBR$^+$11] design an algorithm with similar guarantees for stability to the *min-sum* objective. The authors then apply their algorithms to clustering protein sequences from two large datasets, Pfam and SCOP, comparing the output of their algorithms and existing methods against a 'ground-truth' manual classification. They demonstrate that their accuracy on Pfam is substantially better than, and on SCOP a little better than (but much faster than) the best competing method. Interestingly, on these datasets, the algorithm designed for the k-median objective achieves generally higher accuracy than the algorithm designed for min-sum.

[ABS10] This work considers the promise condition (for k-median or k-means objectives) that merging any two clusters in the optimal solution for that objective raises the cost of that solution by a factor $1+\alpha$ for some constant $\alpha > 0$. This is a relaxation of $(1+\alpha, \epsilon)$-approximation stability in the case that all clusters in the target have size $> \epsilon n$.[3] Under this condition, [ABS10] achieve a PTAS for the given objective (a $1+\delta$ approximation in time polynomial in n and k but exponential in $1/\delta$ and $1/\alpha$).[4] This in turn implies the solution is ϵ-close to the target if approximation-stability indeed holds.

[ABB$^+$10, BB11] This work considers approximation-stability and related conditions for the problem of finding equilibria in 2-player general-sum games. Awasthi et al. [ABB$^+$10] considers the condition that all approximate equilibria are contained within a small ball around

[3] Because in that case, *any* solution using only $k-1$ clusters would have distance greater than ϵ from the target, and approximation-stability would therefore require that all such clusterings have cost greater than $(1+\alpha)\Phi(\mathbf{OPT})$.

[4] Earlier work of Ostrovsky et al. [ORSS06] achieved this guarantee but under the condition that the cost of the optimal $(k-1)$-clustering should be an $\Omega(1/\delta^2)$ factor larger than the cost of the optimal k-clustering.

some true equilibrium (e.g., if the goal is to predict behavior that you believe is at a Nash equilibrium, then this is the condition you would want in order for an approximate equilibrium to be a good prediction). It gives a polynomial-time algorithm to find approximate equilibria if this ball is very small, and more generally gives improved running time over [LMM03] as a function of the ball radius. Balcan and Braverman [BB11] considers a weaker condition, only requiring that for every approximate equilibrium (or even just every well-supported approximate equilibrium) there exist some Nash equilibrium that is close. They also relate this condition to the notion of perturbation-resilience of Bilu and Linial [BL10], as well as connect to related notions considered in [LMM06].

References

[ABB+10] P. Awasthi, M. F. Balcan, A. Blum, O. Sheffet, and S. Vempala. On nash-equilibria of approximation-stable games. In *Proc. 3rd International Symp. Algorithmic Game Theory*, 2010.

[ABS10] P. Awasthi, A. Blum, and O. Sheffet. Stability yields a PTAS for k-median and k-means clustering. In *Proc. 51st Annual IEEE Symp. Foundations of Computer Science (FOCS)*, 2010.

[AGK+04] V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit. Local search heuristics for k-median and facility location problems. *SIAM J. Comput.*, 33(3):544–562, 2004.

[ARV04] S. Arora, S. Rao, and U. Vazirani. Expander flows, geometric embeddings, and graph partitioning. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, 2004.

[BB09] M.-F. Balcan and M. Braverman. Finding low error clusterings. In *COLT*, 2009.

[BB11] M. F. Balcan and M. Braverman. Approximate Nash Equilibria under Stability Conditions. Manuscript, 2011.

[BBG09] M.-F. Balcan, A. Blum, and A. Gupta. Approximate clustering without the approximation. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2009. Long version available at `http://www.cs.cmu.edu/~avrim/Papers/bbg-clustering-2010.pdf`.

[BL10] Y. Bilu and N. Linial. Are stable instances easy? In *Proceedings of the First Symposium on Innovations in Computer Science*, 2010.

[CGTS99] M. Charikar, S. Guha, E. Tardos, and D. B. Shmoy. A constant-factor approximation algorithm for the k-median problem. In *Proc. 31st Annual ACM Symp. Theory of Computing*, 1999.

[JMS02] K. Jain, M. Mahdian, and A. Saberi. A new greedy approach for facility location problems. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, 2002.

[LMM03] Richard J. Lipton, Evangelos Markakis, and Aranyak Mehta. Playing large games using simple strategies. In *Proc. 4th ACM Conf. Electronic Commerce*, pages 36–41, 2003.

[LMM06] R. J. Lipton, E. Markakis, and A. Mehta. On stability properties of economic solution concepts. Manuscript, 2006.

[ORSS06] R. Ostrovsky, Y. Rabani, L. Schulman, and C. Swamy. The effectiveness of Lloyd-type methods for the k-means problem. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, 2006.

[VBR+10] K. Voevodski, M. F. Balcan, H. Roeglin, S-H. Teng, and Y. Xia. Efficient clustering with limited distance information. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, 2010.

[VBR+11] K. Voevodski, M.F. Balcan, H. Roeglin, S-H. Teng, and Y. Xia. Min-sum clustering of protein sequences with limited distance information. In *Proc. of the 1st International Workshop on Similarity-Based Pattern Analysis and Recognition (SIMBAD)*, 2011.