Machine Unlearning via Algorithmic Stability *

Enayat Ullah † Tung Mai‡ Anup Rao § Ryan Rossi¶ Raman Arora ‖

We study the problem of machine unlearning and identify a notion of algorithmic stability, Total Variation (TV) stability, which we argue, is suitable for the goal of *exact* unlearning. For convex risk minimization problems, we design TV-stable algorithms based on noisy Stochastic Gradient Descent (SGD). Our key contribution is the design of corresponding *efficient* unlearning algorithms, which are based on constructing a (maximal) coupling of Markov chains for the noisy SGD procedure. To understand the trade-offs between accuracy and unlearning efficiency, we give upper and lower bounds on excess empirical and populations risk of TV stable algorithms for convex risk minimization. Our techniques generalize to arbitrary non-convex functions, and our algorithms are differentially private as well.

*The full version of the paper can be found at https://arxiv.org/abs/2102.13179

†Johns Hopkins University. Email: enayat@jhu.edu
‡Adobe Research. Email: tumai@adobe.com
§Adobe Research. Email: anuprao@adobe.com
¶Adobe Research. Email: rrossi@adobe.com
‖Johns Hopkins University. Email: arora@cs.jhu.edu

2nd Symposium on Foundations of Responsible Computing (FORC), June 9-11, 2021