Starting Small – Learning with Adaptive Sample Sizes

Hadi Daneshmand Aurélien Lucchi Thomas Hofmann

Institute for Machine Learning, ETH Zürich & Max-Planck ETH Center for Learning Systems

Empirical Risk Minimization

▷ Supervised machine learning as empirical risk minimization

▷ Empirical risk minimizer with i.i.d. sample $\mathcal{S}$, $|\mathcal{S}| := n$

$$\mathcal{R}_\mathcal{S}(\mathbf{w}) := \frac{1}{n}\sum_{\mathrm{x}\in\mathcal{S}} f_\mathrm{x}(\mathbf{w}), \quad \mathbf{w}^*_\mathcal{S} := \operatorname*{argmin}_{\mathrm{w}\in\mathcal{F}} \mathcal{R}_\mathcal{S}(\mathbf{w})\,.$$

▷ Expected error $\le$ estimation error + optimization error (cf. Bottou, Bousquet 2008 [BB08])

$$\mathbf{E}_\mathcal{S}\mathcal{R}(\mathbf{w}^t) - \mathcal{R}^* \le \mathcal{H}(n) + \epsilon(t)\,, \quad \epsilon(t) \sim \text{opt. error}$$

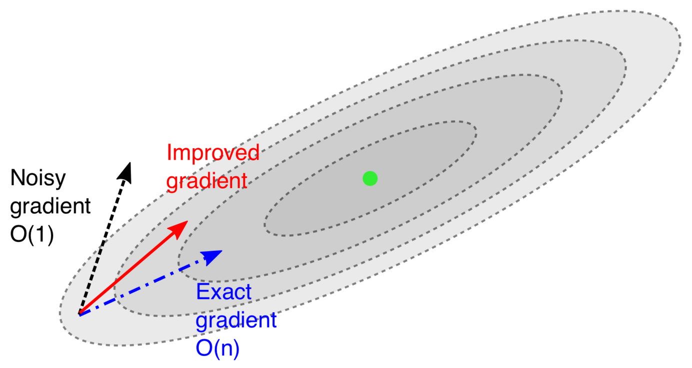
▷ Uniform convergence bounds:

$$\mathbf{E}_\mathcal{S}\left[\mathcal{R}^*_\mathcal{S} - \mathcal{R}^*\right] \le \mathcal{H}(n)\,, \quad \text{e.g. } \mathcal{H}(n) \propto \frac{D}{n}$$

Fast Stochastic Gradient Descent

▷ Variance reduced SGD, e.g. SAGA, SVRG, SAG, etc. (cf. Defazio et al., 2014 [DBLJ14])

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta\left[\nabla f_\mathrm{x}(\mathbf{w}^t) - g_\mathrm{x}\right]$$
$$g_\mathrm{x} := \nabla f_\mathrm{x}(\mathbf{w}^{old}) - \tilde{\nabla}\mathcal{R}_\mathcal{S}$$



▷ Linear convergence: geometric rate

$$\mathbf{E}_\mathcal{A}\left[\mathcal{R}_\mathcal{S}(\mathbf{w}^t) - \mathcal{R}^*_\mathcal{S}\right] \le \rho_n^t C, \quad \rho_n = 1 - \min\left\{\frac{1}{n}, \frac{\mu}{L}\right\}$$

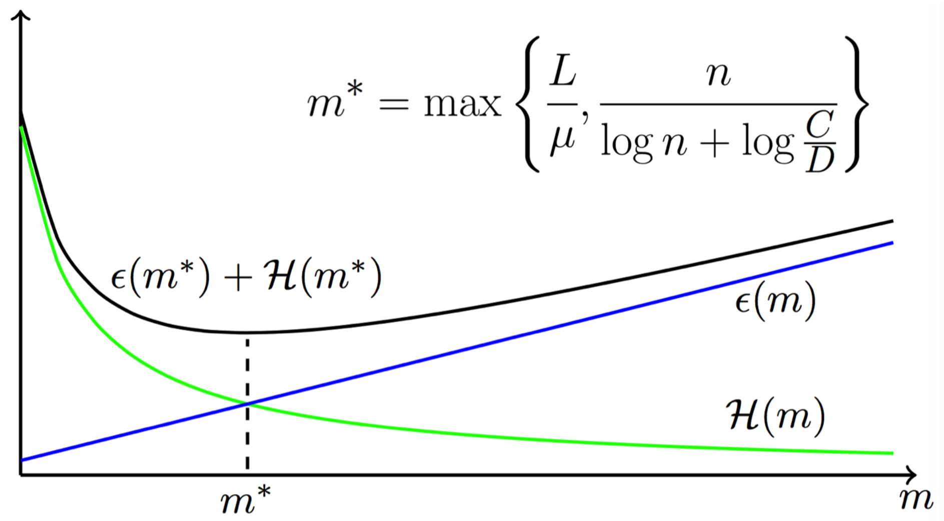
– L: Lipschitz constant (smoothness), μ: strong convexity

– big data limit: dominated by $1/n$

Static Sample Size Optimization

▷ Baseline: one "epoch" with full sample

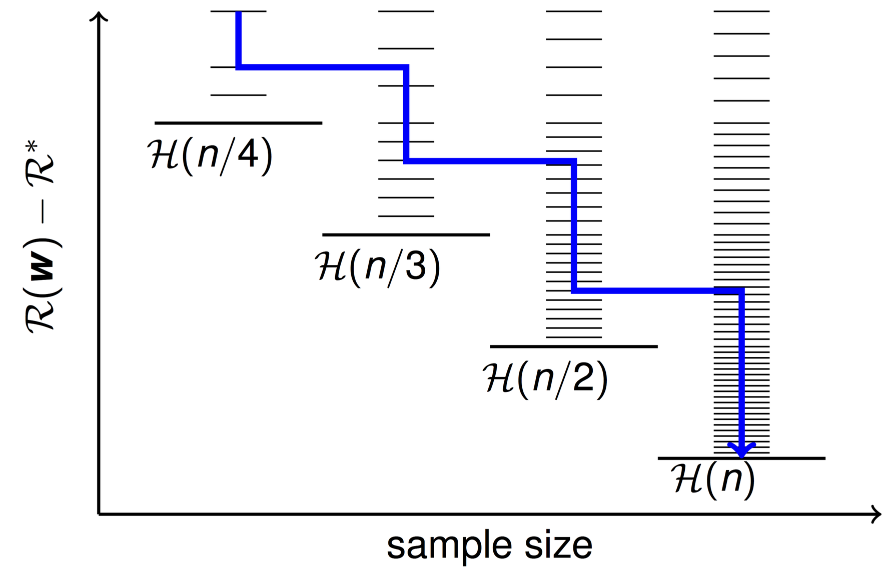
$$\epsilon(t) \le \left(1 - \frac{1}{n}\right)^n C \simeq \frac{C}{e}.$$



Dynamic Sample Size Optimization

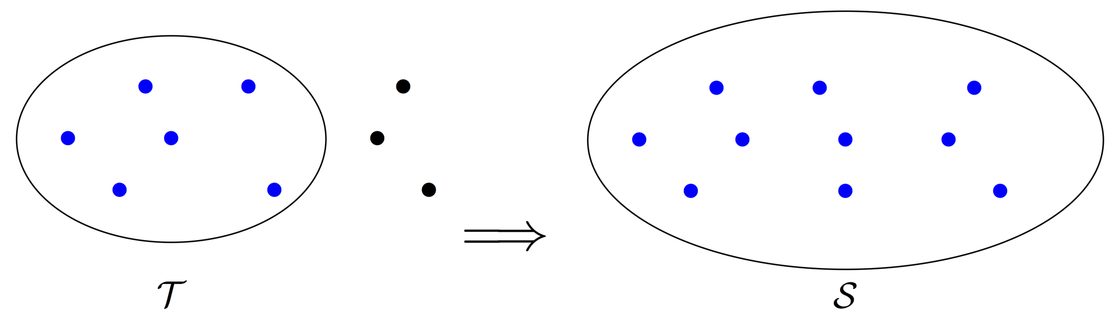
▷ Start with a small sample set to converge faster

▷ Add new samples to lower estimation error



Sample Size Strategy

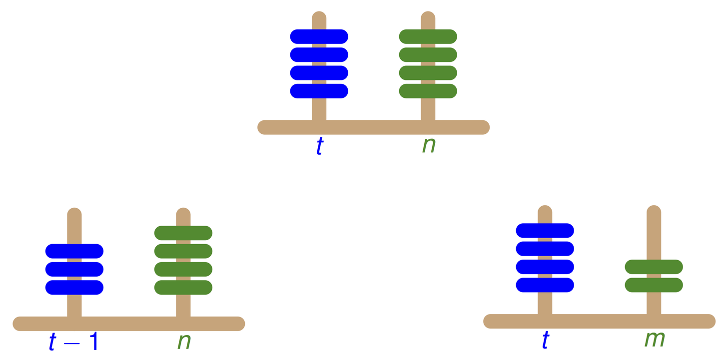
▷ Switching from sample $\mathcal{T}$, $|\mathcal{T}| = m$ to $\mathcal{S} \supseteq \mathcal{T}$, $|\mathcal{S}| = n$



$\mathcal{R}_\mathcal{T}(\boldsymbol{w}) - \mathcal{R}^*_\mathcal{T} \le \epsilon$ $\mathbf{E}_\mathcal{S}\left[\mathcal{R}_\mathcal{S}(\boldsymbol{w}) - \mathcal{R}^*_\mathcal{S}\right] \le \epsilon + \boxed{\frac{n-m}{n}\mathcal{H}(m)}$

▷ Optimal bound/strategy by induction

$$\mathbf{U}(t,n) = \min\begin{cases}\rho_n \mathbf{U}(t-1,n)\\ \min\limits_{m<n}\left[\mathbf{U}(t,m) + \frac{n-m}{n}\mathcal{H}(m)\right]\end{cases}$$



▷ Suggested schedule for $\mathbf{U}(n,n)$: $m(t) = \max\left\{2\frac{L}{\mu}, \left\lceil\frac{t}{2}\right\rceil\right\}$ iterate 2, add 1 (variant: always iterate on new sample once)

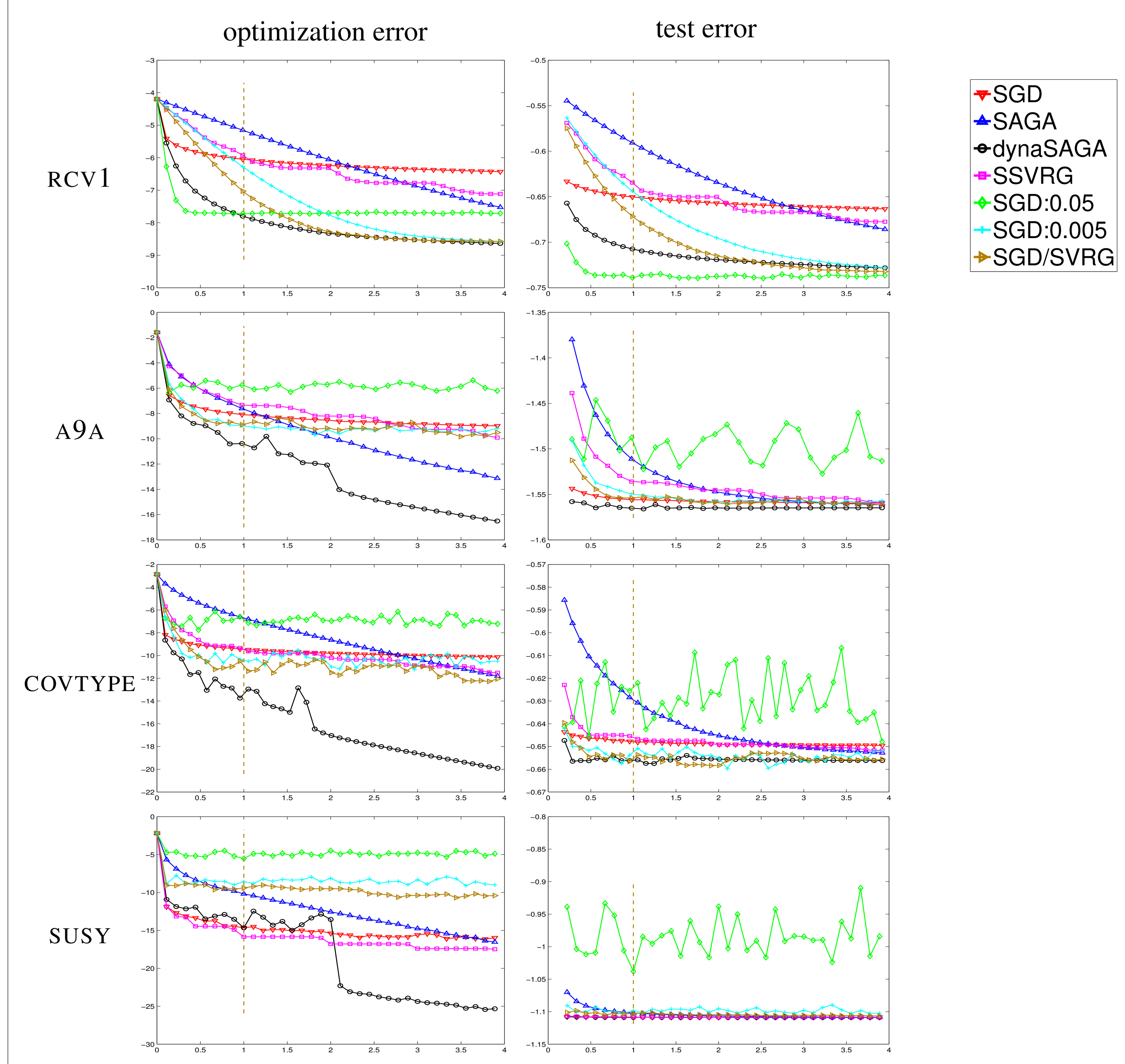
Analysis

▷ **Lemma**: For $\mathcal{H}(n) \propto D/n$, the "$m = 2t$"-strategy minimizes $\mathbf{U}(n,n)$ for all sample sizes $n > \kappa$, $\kappa = L/\mu$.

▷ **Lemma & Corollary**: For $\mathcal{H}(n) = Dn^{-\alpha}$, $\alpha \le 1$, it holds that:

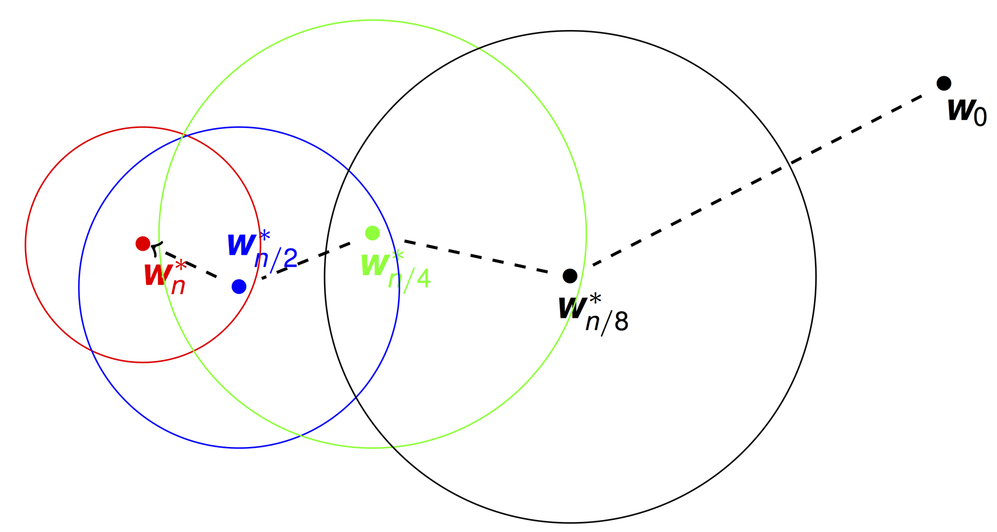
$$\mathbf{U}(n,n) \le \left(3\cdot 2^{\alpha-1}\right)\mathcal{H}(n) + 2\xi\left(\frac{\kappa}{n}\right)^2$$

Experiments



Future Work

▷ DYNANEWTON (cf. Daneshmand et al., 2014 [DLH16])



References

[BB08] Olivier Bousquet and Léon Bottou. The tradeoffs of large scale learning. In *NIPS*, 2008.

[DBLJ14] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, 2014.

[DLH16] Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Dynanewton-accelerating newton's method for machine learning. *arXiv*, 2016.