An Implicit Form of Krasulina's k-PCA Update without the Orthonormality Constraint

Ehsan Amid and Manfred K. Warmuth
University of California, Santa Cruz
Google Brain, Mountain View
{eamid, manfred}@google.com

Implicit vs. Explicit Update

Gradient descent is motivated by

$$\theta_{t+1} = \underset{\theta}{\operatorname{argmin}} \left({}^1\!/_{2\eta} \underbrace{\|\theta - \theta_t\|^2}_{\text{inertia}} + \text{loss}(\theta) \right)$$

The actual minimizer of **inertia + loss**

$$\theta_{t+1} = \theta_t - \eta\, \nabla \text{loss}(\theta_{t+1}) \qquad \textbf{(implicit update)}$$

Commonly approximated by the old gradient

$$\theta_{t+1} \approx \theta_t - \eta\, \nabla \text{loss}(\theta_t) \qquad \textbf{(explicit update)}$$

Most of Machine Learning is explicit!

Implicit Krasulina Update

Without the orthonormality constraint

$$\mathrm{C}^{\text{new}} = \underset{\widetilde{\mathrm{C}}}{\operatorname{argmin}}\ {}^1\!/_2 \left({}^1\!/_\eta \|\widetilde{\mathrm{C}} - \mathrm{C}\|_{\mathrm{F}}^2 + \mathbb{E}\left[\|\widetilde{\mathrm{C}}\, \widetilde{\mathrm{C}}^\dagger \mathrm{y} - \mathrm{y}\|^2\right] \right)$$

Approximating $\widetilde{\mathrm{C}}^\dagger \mathrm{y}$ with $\mathrm{C}^\dagger \mathrm{y}$ yields the (partially) **Implicit Krasulina (a.k.a. Sanger)** update:

$$\mathrm{x}_t = \mathrm{C}^\dagger \mathrm{y}_t \qquad \text{(E-Step)}$$

$$\mathrm{C}^{\text{new}} = \mathrm{C} - \frac{\eta}{1 + \eta\, \|\mathrm{x}_t\|^2} (\mathrm{C}\mathrm{x}_t - \mathrm{y}_t)\, \mathrm{x}_t^\top \qquad \text{(M-Step)}$$

- No QR-step, needs to keep track of $\mathrm{C}^\dagger$ instead
- Has an **online EM** interpretation! (see [1])

k-PCA Loss

k-PCA: Given zero mean random variable $\mathrm{y} \in \mathbb{R}^d$
Find the (d x d) projection matrix P of rank-k s.t.

$$\ell_{\text{comp}}(\mathrm{P}) = \mathbb{E}\left[\|\mathrm{P}\,\mathrm{y} - \mathrm{y}\|^2\right] = \operatorname{tr}\left((\mathrm{I}_d - \mathrm{P})\, \mathbb{E}[\mathrm{y}\,\mathrm{y}^\top]\right)$$

or

$$\ell_{\text{var}}(\mathrm{P}) = -\mathrm{P}\, \operatorname{tr}(\mathbb{E}[\mathrm{y}\mathrm{y}^\top])$$

is minimized.

Online k-PCA: Observe one example y_t at a time

Decomposition: $\mathrm{P} = \mathrm{C}\,(\mathrm{C}^\top \mathrm{C})^{-1} \mathrm{C}^\top := \mathrm{C}\mathrm{C}^\dagger$

Solve for C instead of P!

GD on the Stiefel Manifold

Manifold of (d x k) orthonormal matrices

$$\mathrm{St}_{(d,k)} = \{\mathrm{C} \in \mathbb{R}^{d\times k} \,|\, \mathrm{C}^\top \mathrm{C} = \mathrm{I}_k\}$$

Krasulina's update: uses <u>projected</u> gradient

$$\widetilde{\mathrm{C}} = \mathrm{C} - \eta\, \overline{\nabla}\, \hat{\ell}_{\text{comp}}(\mathrm{C}) = \mathrm{C} - \eta\, (\mathrm{C}\,\mathrm{x}_t - \mathrm{y}_t)\, \mathrm{x}_t^\top$$

$$\text{where} \quad \mathrm{x}_t = \mathrm{C}^\top \mathrm{y}_t \quad \text{and} \quad \mathrm{C}^{\text{new}} = \mathcal{QR}(\widetilde{\mathrm{C}})$$

Oja's update: uses <u>unprojected</u> gradient

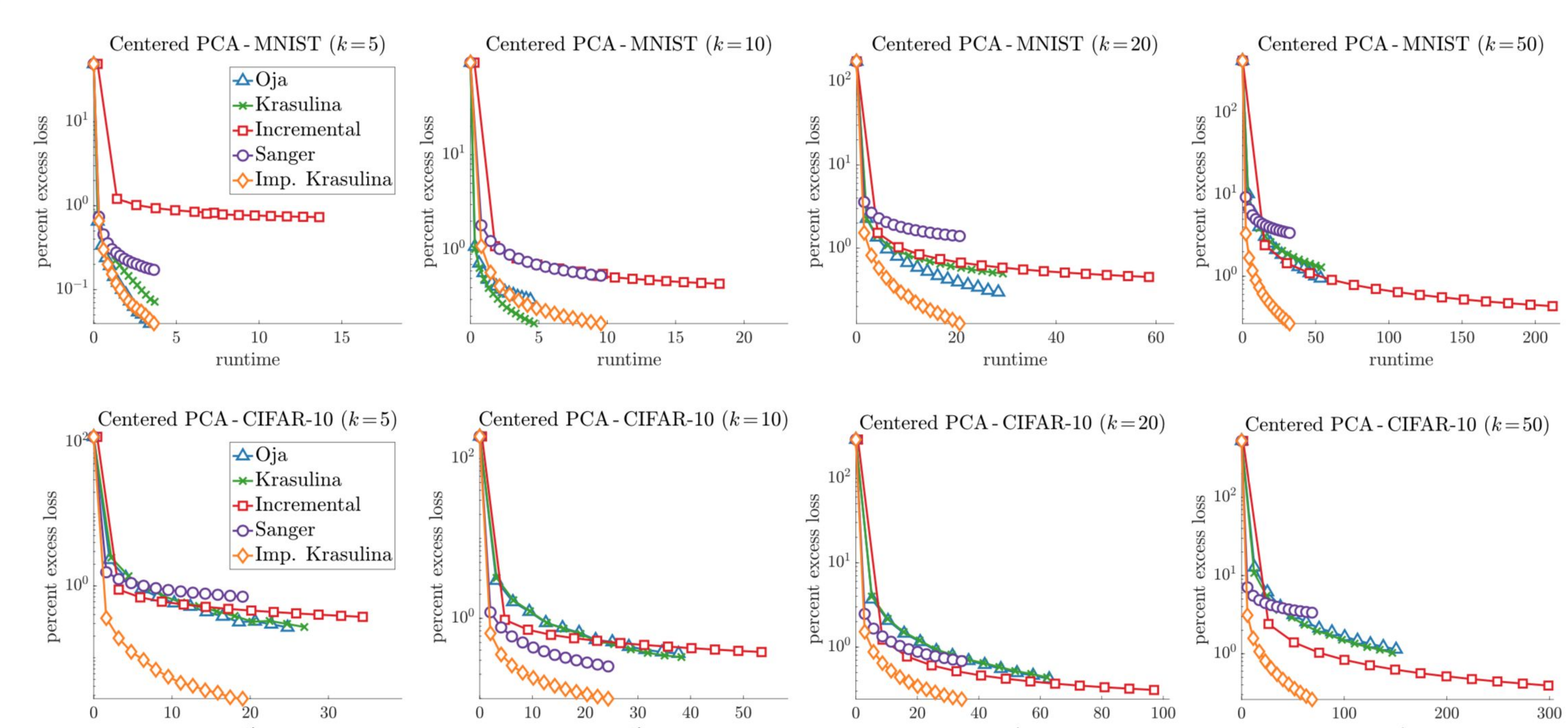
$$\widetilde{\mathrm{C}} = \mathrm{C} - \eta \nabla \hat{\ell}_{\text{var}}(\mathrm{C}) = \mathrm{C} + \eta\, \mathrm{y}_t \mathrm{y}_t^\top\, \mathrm{C}$$

$$\text{and} \quad \mathrm{C}^{\text{new}} = \mathcal{QR}(\widetilde{\mathrm{C}})$$

Both updates are explicit!

Results

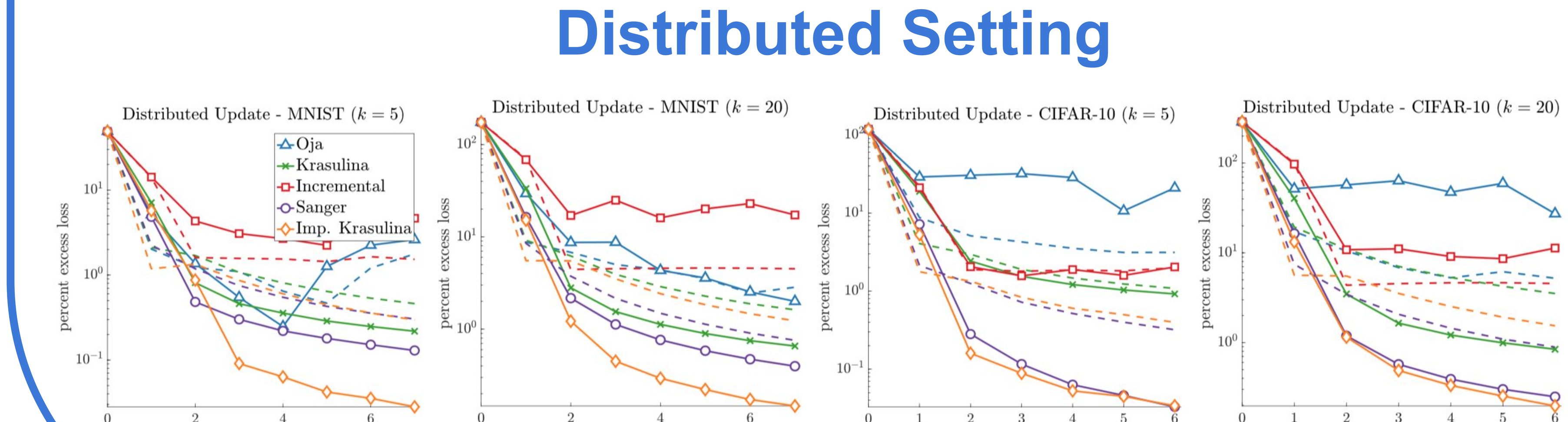
Online k-PCA



Sensitivity to Learning Rate

Method	η_0-Scale	MNIST			CIFAR10		
		$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$
Batch PCA	–	35.16	26.95	18.74	86.98	65.70	48.65
Oja	0.1×	35.79 ± 0.38	29.78 ± 0.38	21.55 ± 0.25	116.26 ± 2.58	66.81 ± 0.58	58.28 ± 0.65
	1×	35.19 ± 0.05	26.98 ± 0.02	18.80 ± 0.03	87.28 ± 0.43	65.90 ± 0.01	48.86 ± 0.05
	10×	35.29 ± 0.01	27.08 ± 0.01	19.01 ± 0.01	87.22 ± 0.04	67.68 ± 0.13	50.36 ± 0.13
Krasulina	0.1×	35.78 ± 0.37	29.92 ± 0.37	21.44 ± 0.24	116.86 ± 2.58	66.76 ± 0.65	57.62 ± 0.65
	1×	35.17 ± 0.00	26.96 ± 0.01	18.79 ± 0.02	87.04 ± 0.52	65.90 ± 0.01	48.87 ± 0.05
	10×	35.30 ± 0.01	27.09 ± 0.01	19.02 ± 0.02	87.17 ± 0.04	67.87 ± 0.16	50.56 ± 0.15
Incremental	–	35.26 ± 0.10	27.01 ± 0.04	18.83 ± 0.05	87.50 ± 0.46	65.75 ± 0.05	48.82 ± 0.10
Sanger	0.1×	39.09 ± 0.78	30.61 ± 0.48	22.03 ± 0.23	90.74 ± 1.84	70.70 ± 2.18	54.04 ± 0.95
	1×	36.27 ± 0.93	28.00 ± 0.38	19.82 ± 0.18	88.21 ± 0.81	66.84 ± 0.88	50.73 ± 0.62
	10×	42.36 ± 1.62	34.48 ± 0.95	25.50 ± 0.46	99.78 ± 3.69	77.88 ± 2.62	60.21 ± 1.08
Imp. Krasulina	0.1×	35.17 ± 0.01	26.96 ± 0.01	18.78 ± 0.02	87.00 ± 0.00	65.75 ± 0.05	48.75 ± 0.03
	1×	35.17 ± 0.01	26.97 ± 0.03	18.77 ± 0.02	87.01 ± 0.00	65.78 ± 0.09	48.74 ± 0.04
	10×	35.17 ± 0.01	26.98 ± 0.05	18.77 ± 0.01	87.02 ± 0.04	65.74 ± 0.01	48.76 ± 0.03

Distributed Setting



Reference: [1] E. Amid and M. K. Warmuth, Divergence-based Motivation for Online EM and Combining Hidden Variable Models, *arXiv preprint arXiv:1902.04107*, 2019.