Weierstrass Institute for
Applied Analysis and Stochastics

Leibniz Association

On a Combination of Alternating Minimization and Nesterov's Momentum

Sergey Guminov, **Pavel Dvurechensky**, Nazarii Tupitsa, Alexander Gasnikov

International Conference on Machine Learning 2021

- $\hat{r}_{ui}$ – unknown ratings for user u and item i
- Structural assumption: $\hat{r}_{ui} = {x_u}^\top y_i +$ noise
- Partial observations r_{ui}

x and y are found from optimization problem:

$$\min_{x,y} \; F(x,y) = \sum_{\text{observed } u,i} c_{ui} \left(r_{ui} - {x_u}^\top y_i\right)^2 + \alpha \sum_u \|x_u\|_2^2 + \alpha \sum_i \|y_i\|_2^2.$$

- Easy alternating minimization (AM):
  explicit minimization in x for fixed y and vice versa.
- Non-convex problem.
- Two blocks of variables.

Given two histograms $a, b \in S_n(1)$ and a cost matrix $C \in \mathbb{R}_+^{n\times n}$ solve the entropy-regularized optimal transport (OT) problem:

Primal problem: $$\min_{X\in\mathcal{U}(a,b)} \langle C, X\rangle + \gamma\langle X, \ln X\rangle,$$

$$\mathcal{U}(a,b) := \{X \in \mathbb{R}_+^{n\times n} : \; X\mathbf{1} = a, \; X^T\mathbf{1} = b\}$$

Dual problem: $$\min_{u,v\in\mathbb{R}^n} \left\{\ln\left(\left(e^{u}\right)^\top e^{-C/\gamma} e^{v}\right) - \langle u, a\rangle - \langle v, b\rangle\right\}$$

- Strongly convex linearly-constrained primal, smooth convex dual.
- Easy AM in the dual: explicit minimization in u for fixed v, two blocks.
- Sinkorn's algorithm as AM in the dual, rate is proportional to $1/k$.
- Accelerated Gradient Method (AGM) in the dual has rate proportional to $1/k^2$.
- Primal-dual analysis needed to find the primal variable X.

Can we combine Alternating Minimization and Nesterov's (Momentum) Accelerated Gradient Method?

Desired properties: analysis for non-convex problems (Collaborative Filtering), primal-dual analysis (Entropic Optimal Transport), many blocks, parameter-free: no need to know Lipschitz constant, etc.

	P-F	Acc.	N-C	P-D	B-N
AM (Beck & Tetruashvili, 2013; Beck, 2015)	$\surd$	$\times$	$\times$	$\times$	2
AM (Saha & Tewari, 2013; Sun & Hong, 2015)	$\surd$	$\times$	$\times$	$\times$	any
ACD (Nesterov, 2012; Lee & Sidford, 2013; Fercoq & Richtárik,2015) and many others	$\times$	$\surd$	$\times$	$\surd$	any
AAR-BCD (Diakonikolas & Orecchia, 2018)	$\times$	$\surd$	$\times$	$\times$	any
AAM (Diakonikolas & Orecchia, 2018)	$\surd$	$\surd$	$\times$	$\times$	2
This paper	$\surd$	$\surd$	$\surd$	$\surd$	any

$$\min_{\lambda \in \mathbb{R}^N} \varphi(\lambda).$$

Examples of φ:

- non-convex objective in collaborative filtering;
- convex dual objective in the entropic OT.

Definitions and assumptions:

- n blocks $I_p, p \in \{1, \ldots, n\}$.
- Block minimization: we can solve

$$\min_{\lambda} \left\{ \varphi(\lambda) : \lambda \in S_p(\zeta) := \{\zeta + \operatorname{span}\{e_i : \ i \in I_p\}\} \right\}$$

- $\varphi(\lambda)$ is L_φ-smooth: $\forall \ \lambda, \ \eta \in \mathbb{R}^N \quad \|\nabla \varphi(\lambda) - \nabla \varphi(\eta)\|_2 \leqslant L_\varphi \|\lambda - \eta\|_2.$
- Notation: λ_i – components of λ corresponding to the block i and $\nabla_i \varphi(\lambda)$ – gradient corresponding to the block i.

1: $\beta_0 = \alpha_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$.
2: **for** $k \geqslant 0$ **do**
3: [Coupling step]

Set $\lambda_k = \tau_k \zeta_k + (1 - \tau_k)\eta_k$, $\tau_k = \arg\min_{\tau \in [0,1]} \varphi\left(\eta_k + \tau(\zeta_k - \eta_k)\right)$

4: [Gauss-Southwell] Choose $i_k = \arg\max_{i \in \{1,\dots,n\}} \|\nabla_i \varphi(\lambda_k)\|_2^2$.
5: [Block Minimization] Set $\eta_{k+1} = \arg\min_{\eta \in S_{i_k}(\lambda_k)} \varphi(\eta)$. (instead of gradient step)
6: Find α_{k+1}, $\beta_{k+1} := \beta_k + \alpha_{k+1}$ from

$$\varphi(\lambda_k) - \frac{\alpha_{k+1}^2}{2(\beta_k + \alpha_{k+1})} \|\nabla \varphi(\lambda_k)\|_2^2 = \varphi(\eta_{k+1})$$

7: [Update momentum] Set $\zeta_{k+1} = \zeta_k - \alpha_{k+1} \nabla \varphi(\lambda_k)$.
8: [Optional primal update] Set $\hat{x}_{k+1} = \frac{\alpha_{k+1} x(\lambda_k) + \beta_k \hat{x}_k}{\beta_{k+1}}$.
9: **end for**

Ensure: η_{k+1} and optional $\hat{x}_{k+1}$.

$\beta_0 = \alpha_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$.

for $k \geqslant 0$ **do**

  [Coupling step] Set $\lambda_k = \tau_k \zeta_k + (1 - \tau_k)\eta_k$,
  $\tau_k = \arg\min_{\tau \in [0,1]} \varphi\left(\eta_k + \tau(\zeta_k - \eta_k)\right)$

  [Gauss-Southwell] Choose $i_k = \arg\max_{i \in \{1,\dots,n\}} \|\nabla_i \varphi(\lambda_k)\|_2^2$.

  [Block Minimization]

Set $\eta_{k+1} = \arg\min_{\eta \in S_{i_k}(\lambda_k)} \varphi(\eta)$. (instead of gradient step)

  Find α_{k+1}, $\beta_{k+1} := \beta_k + \alpha_{k+1}$ from

$$\varphi(\lambda_k) - \frac{\alpha_{k+1}^2}{2(\beta_k + \alpha_{k+1})} \|\nabla \varphi(\lambda_k)\|_2^2 = \varphi(\eta_{k+1})$$

  [Update momentum] Set $\zeta_{k+1} = \zeta_k - \alpha_{k+1} \nabla \varphi(\lambda_k)$.

  [Optional primal update] Set $\hat{x}_{k+1} = \frac{\alpha_{k+1} x(\lambda_k) + \beta_k \hat{x}_k}{\beta_{k+1}}$.

end for

Ensure: η_{k+1} and optional $\hat{x}_{k+1}$.

$\beta_0 = \alpha_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$.

for $k \geqslant 0$ **do**

[Coupling step] Set $\lambda_k = \tau_k \zeta_k + (1 - \tau_k)\eta_k$,

$\tau_k = \arg \min\limits_{\tau \in [0,1]} \varphi\left(\eta_k + \tau(\zeta_k - \eta_k)\right)$

[Gauss-Southwell] Choose $i_k = \arg \max\limits_{i \in \{1,\dots,n\}} \|\nabla_i \varphi(\lambda_k)\|_2^2$.

[Block Minimization] Set $\eta_{k+1} = \arg \min\limits_{\eta \in S_{i_k}(\lambda_k)} \varphi(\eta)$. (instead of gradient step)

Find α_{k+1}, $\beta_{k+1} := \beta_k + \alpha_{k+1}$ from

$$\varphi(\lambda_k) - \frac{\alpha_{k+1}^2}{2(\beta_k + \alpha_{k+1})} \|\nabla \varphi(\lambda_k)\|_2^2 = \varphi(\eta_{k+1})$$

[Update momentum]

Set $\zeta_{k+1} = \zeta_k - \alpha_{k+1} \nabla \varphi(\lambda_k)$.

[Optional primal update] Set $\hat{x}_{k+1} = \frac{\alpha_{k+1} x(\lambda_k) + \beta_k \hat{x}_k}{\beta_{k+1}}$.

end for

Ensure: η_{k+1} and optional $\hat{x}_{k+1}$.

$\beta_0 = \alpha_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$.

for $k \geqslant 0$ **do**

[Coupling step] Set $\lambda_k = \tau_k \zeta_k + (1 - \tau_k)\eta_k$,
$\tau_k = \arg\min_{\tau \in [0,1]} \varphi\left(\eta_k + \tau(\zeta_k - \eta_k)\right)$

[Gauss-Southwell] Choose $i_k = \arg\max_{i \in \{1,\dots,n\}} \|\nabla_i \varphi(\lambda_k)\|_2^2$.

[Block Minimization] Set $\eta_{k+1} = \arg\min_{\eta \in S_{i_k}(\lambda_k)} \varphi(\eta)$. (instead of gradient step)

Find α_{k+1}, $\beta_{k+1} := \beta_k + \alpha_{k+1}$ from

$$\varphi(\lambda_k) - \frac{\alpha_{k+1}^2}{2(\beta_k + \alpha_{k+1})}\|\nabla \varphi(\lambda_k)\|_2^2 = \varphi(\eta_{k+1})$$

[Update momentum] Set $\zeta_{k+1} = \zeta_k - \alpha_{k+1}\nabla\varphi(\lambda_k)$.

[Optional primal update]

Set $\hat{x}_{k+1} = \frac{\alpha_{k+1} x(\lambda_k) + \beta_k \hat{x}_k}{\beta_{k+1}}$.

end for

Ensure: η_{k+1} and optional $\hat{x}_{k+1}$.

Convex setting:

$$\varphi(\eta_k) - \varphi(\lambda^*) \leq \frac{2nL_\varphi \|\lambda_0 - \lambda^*\|_2^2}{k^2} = O\left(\frac{n}{k^2}\right),$$

Non-convex setting:

$$\min_{i=0,\dots,k} \|\nabla\varphi(\lambda_i)\|_2^2 \leqslant \frac{2nL_\varphi(\varphi(\lambda_0) - \varphi(\lambda^*))}{k} = O\left(\frac{n}{k}\right).$$

Automatic adaptation to convexity and smoothness constant.

Prinal-dual setting:

Primal objective f is γ-strongly convex, dual solution $\|\lambda^*\|_2 \leq R$.

$$f(\hat{x}_k) - f^* \leq f(\hat{x}_k) + \varphi(\eta_k) \leq \frac{4nL_\varphi R^2}{k^2} = \frac{8n\|A\|_{E\to H}^2 R^2}{\gamma k^2} = O\left(\frac{n}{\gamma k^2}\right),$$

$$\|A\hat{x}_k - b\|_2 \leq \frac{8n\|A\|_{E\to H}^2 R}{\gamma k^2} = O\left(\frac{n}{\gamma k^2}\right),$$

Thank you!

More details in the paper:

- Numerical experiments
- State-of-the-art complexity for Optimal Transport and Barycenters