Smooth Bilevel Programming for Sparse Regularization

Gabriel Peyré

Joint work with:
Clarice Poon
(Bath University)
Iterative Reweighted Least squares

Consider \[ \min_k \left( \sum_i x_i^k y_k \right) \]

Quadratic variational formulation:

\[ \min_{\| \cdot \|_2} \min_\tau \left( \sum_i \tau_i \right) \]

Regularise with \( \tau > 0 \):

\[ \min_{\| \cdot \|_2} \min_\tau \left( \sum_i \tau_i \right) \]

Alternating minimisation:

\[ \begin{align*}
X_{k+1} &= X_k + \text{diag}(1/\tau_k) (1 + 1) \\
\tau_{k+1} &= \text{quad-variational} \quad \text{Non-cvx.-Alternating-min} \quad \text{Non-cvx.-LBFGS} \quad \text{CELER} \quad \text{Noncvx-Pro}
\end{align*} \]

\[ \begin{align*}
\beta &= 0.2 \\
\beta &= 0.6 \\
\beta &= 1.0
\end{align*} \]
Sparse Regularization

\[ A \in \mathbb{R}^{n \times d} \quad Ax \approx y \quad \text{over-determined + noise} \quad \text{under-determined} \quad \text{regularize} \]

\[
\min_{x \in \mathbb{R}^d} \frac{1}{2\lambda} \|Ax - y\|^2 + \|x\|_1 \quad \lambda \to 0 \quad \min_{Ax=y} \|x\|_1
\]
Sparse Regularization

\[ A \in \mathbb{R}^{n \times d} \quad Ax \approx y \]

over-determined + noise
under-determined

\[ \min_{x \in \mathbb{R}^d} \frac{1}{2\lambda} \|Ax - y\|^2 + \|x\|_1 \]

\[ \lambda \to 0 \]

\[ \min_{Ax=y} \|x\|_1 \]

Spike deconvolution: \( Ax = a \ast x \)
Sparse Regularization

\[ A \in \mathbb{R}^{n \times d} \quad A x \approx y \]

\[
\min_{x \in \mathbb{R}^d} \frac{1}{2\lambda} \|Ax - y\|^2 + \|x\|_1
\]

\[
\lambda \to 0 \quad \min \|x\|_1
\]

over-determined + noise
under-determined

regularize

Spike deconvolution: \( Ax = a \ast x \)

Model selection: \( y_i \approx \langle a_i, x \rangle \)

Data \((a_i, y_i)^n_{i=1}\)

\[
\text{Ridge: } \|x\|_2^2
\]

\[
\text{Lasso: } \|x\|_1
\]

\[ y = x \ast a + \varepsilon \]

\[ x_0 \]

\[ x_\lambda \]
Forward-Backward

\[
\min_{x \in \mathbb{R}^d} f(x) \triangleq \frac{1}{2\lambda} \|Ax - y\|^2 + \|x\|_1
\]  

\[
x_{k+1} = \text{Prox}_{\tau\lambda\|\cdot\|_1}(y - \tau A^\top (Ax_k - y))
\]

\[
\text{Prox}_{\sigma\|\cdot\|_1}(x) = (\text{sign}(x_i)(|x_i| - \sigma)_+)i
\]
Moreau-Yosida regularisation

\[
\begin{align*}
\min_{x \in \mathbb{R}^d} f(x) & \triangleq \frac{1}{2\lambda} \|Ax - y\|^2 + \|x\|_1 \\
x_{k+1} &= \text{Prox}_{\tau\lambda\|\cdot\|_1}(y - \tau A^\top(Ax_k - y)) \\
\text{Prox}_{\sigma\|\cdot\|_1}(x) &= (\text{sign}(x_i)(|x_i| - \sigma)_+)i
\end{align*}
\]

\[
\tau < \frac{2}{\|A\|}
\]

Theorem: \( f(x_k) - \min f \leq \frac{C_{n,d}}{k} \)

Nesterov, FISTA accelerations: \( 1/k^2 \) rates.
Forward-Backward

\[
\min_{x \in \mathbb{R}^d} f(x) \triangleq \frac{1}{2\lambda} \|Ax - y\|^2 + \|x\|_1
\]

\[
x_{k+1} = \text{Prox}_{\lambda\|\cdot\|_1}(y - \tau A^\top (Ax_k - y))
\]

\[
\text{Prox}_{\sigma\|\cdot\|_1}(x) = (\text{sign}(x_i)(|x_i| - \sigma)_+)_i
\]

**Theorem:** \( f(x_k) - \min f \leq \frac{C_{n,d}}{k} \)

Nesterov, FISTA accelerations: \(1/k^2\) rates.

Grid-free rates [Léger, Chizat]:
(deconvolution + min-separation)

\[
f(x_k) - \min f \leq \frac{C}{k^{4+\dim}}
\]
Iterative Reweighted Least Squares

“η-trick”

\[ |x| = \inf_{\eta > 0} \left\{ \frac{1}{2} \frac{x^2}{\eta} + \frac{1}{2} \eta \right\} \]

\[ \eta^* = |x| \]
Iterative Reweighted Least Squares

“η-trick” \[ |x| = \inf_{\eta > 0} \frac{1}{2} \frac{x^2}{\eta} + \frac{1}{2} \eta \]

\[
\min_{x \in \mathbb{R}^d} \frac{1}{2\lambda} \|Ax - y\|^2 + \|x\|_1 \quad \text{over parameterization}
\]

\[
\min_{x \in \mathbb{R}^d, \eta \in \mathbb{R}_+} \frac{1}{2\lambda} \|Ax - y\|^2 + \frac{1}{2} \sum_i \frac{x_i^2}{\eta_i} + \eta_i
\]

\[ \eta^* = |x| \]

Jointly convex
Iterative Reweighted Least Squares

"η-trick" \[ |x| = \inf_{\eta > 0} \frac{1}{2} \frac{x^2}{\eta} + \frac{1}{2} \eta \]

Jointly convex over parameterization

Minimization on \( \eta \): \( \eta \leftarrow |x| \)

Minimization on \( x \): \( x \leftarrow (A^\top A + \lambda \text{diag}(\eta^{-1}))^{-1} A^\top y \)
Iterative Reweighted Least Squares

“\( \eta \)-trick” \( |x| = \inf_{\eta > 0} \frac{1}{2} \frac{x^2}{\eta} + \frac{1}{2} \eta \)

\[
\min_{x \in \mathbb{R}^d} \frac{1}{2\lambda} \|Ax - y\|^2 + \|x\|_1
\]

Minimization on \( \eta \): \( \eta \leftarrow |x| \)

Minimization on \( x \): \( x \leftarrow (A^\top A + \lambda \text{diag}(\eta^{-1}))^{-1} A^\top y \)

\[\eta^* = |x|\]

Jointly convex

Problem: does not converge \( (x_i^2 / \eta_i \) non smooth)

Regularisation:

\[
\min_{x \in \mathbb{R}^d, \eta \in \mathbb{R}^d_+} \frac{1}{2\lambda} \|Ax - y\|^2 + \frac{1}{2} \sum_i \frac{x_i^2}{\eta_i} + \eta_i + \varepsilon \eta_i^{-1}
\]

Minimization on \( \eta \): \( \eta \leftarrow \sqrt{|x|^2 + \varepsilon^2} \)
Variational Representation

\[
|x| = \inf_{\eta > 0} \frac{1}{2} \frac{x^2}{\eta} + \frac{1}{2} \eta \\
\eta^* = |x|
\]

\[
u \triangleq \sqrt{\eta} \\
v \triangleq x / \sqrt{\eta}
\]

\[
|x| = \min_{uv = x} \frac{u^2 + v^2}{2}
\]

\[
|u^*| = \text{sign}(x) \sqrt{|x|}
\]
Variational Representation

\[
\begin{align*}
|x| &= \inf_{\eta > 0} \frac{1}{2} \frac{x^2}{\eta} + \frac{1}{2} \eta \\
\eta^* &= \frac{|x|}{2}
\end{align*}
\]

convex non-smooth

\[
|\eta| = \inf_{\eta > 0} \frac{1}{2} \frac{x^2}{\eta} + \frac{1}{2} \eta
\]

\[
\eta^* = \frac{|x|}{2}
\]

\[
\begin{align*}
u &\triangleq \sqrt{\eta} \\
v &\triangleq x / \sqrt{\eta}
\end{align*}
\]

\[
\begin{align*}
|x| &= \min_{uv = x} \frac{u^2 + v^2}{2} \\
u^* &= \text{sign}(x) v^* = \sqrt{|x|}
\end{align*}
\]

non-convex smooth

\[
\begin{align*}
|\eta| &= \inf_{\eta > 0} \frac{1}{2} \frac{x^2}{\eta} + \frac{1}{2} \eta \\
\eta^* &= \frac{|x|}{2}
\end{align*}
\]

Alternate form:

\[
|x| = \min_{u^2 - v^2 = x} u^2 + v^2
\]

→ not amenable to alternate minimization.
Variational Representation

**convex** non-smooth

\[ |x| = \inf_{\eta>0} \frac{1}{2} \frac{x^2}{\eta} + \frac{1}{2} \eta \]

\[ \eta^* = |x| \]

\[ u \triangleq \sqrt{\eta} \]

\[ v \triangleq x/\sqrt{\eta} \]

\[ |x| = \min_{uv=x} \frac{u^2 + v^2}{2} \]

\[ u^* = \text{sign}(x) \]

\[ v^* = \sqrt{|x|} \]

Alternate form:

\[ |x| = \min_{u^2-v^2=x} u^2 + v^2 \]

→ not amenable to alternate minimization.

**non-convex** smooth

\[ x = u \odot v \]

\[ \min_{x \in \mathbb{R}^d} \frac{1}{2\lambda} \|Ax - y\|^2 + \|x\|_1 \]

\[ \|x\|_1 = \sum_{i=1}^d |x_i| \]

\[ \|Ax - y\|_2 = \sqrt{\sum_{i=1}^d (Ax_i - y_i)^2} \]

\[ \Delta = (u_i v_i)_{i=1}^d \]

\[ \min_{u,v} \frac{1}{\lambda} \|A(u \odot v) - y\|^2 + \|u\|_2^2 + \|v\|_2^2 \]
Variational Representation

convex non-smooth

\[ |x| = \inf_{\eta > 0} \frac{1}{2} x^2 \frac{1}{\eta} + \frac{1}{2} \eta \]

\[ \eta^* = |x| \]

\[ u \triangleq \sqrt{\eta} \]

\[ v \triangleq x / \sqrt{\eta} \]

non-convex smooth

\[ |x| = \min_{uv = x} \frac{u^2 + v^2}{2} \]

\[ u^* = \text{sign}(x) v^* = \sqrt{|x|} \]

Alternate form:

\[ |x| = \min_{u^2 - v^2 = x} u^2 + v^2 \]

→ not amenable to alternate minimization.

\[ \min_{x \in \mathbb{R}^d} \frac{1}{2\lambda} \|Ax - y\|^2 + \|x\|_1 \]

\[ x = u \odot v \]

\[ \triangleq (u_i v_i)_{i=1}^d \]

\[ \min_{u,v} \frac{1}{\lambda} \|A(u \odot v) - y\|^2 + \|u\|_2^2 + \|v\|_2^2 \]

**Alternating minimization:**

\[ u \leftarrow (\text{diag}(v) A^\top \text{Adiag}(v) + \lambda \text{Id}_d)^{-1} (v \odot A^\top y) \]

\[ = v \odot A^\top (\text{Adiag}(v^2) A^\top + \lambda \text{Id}_n)^{-1} y \]

\[ v \leftarrow (\text{diag}(u) A^\top \text{Adiag}(u) + \lambda \text{Id}_d)^{-1} (u \odot A^\top y) u \]

(Sherman–Morrison formula)

**Empirical finding:** works better than IRLS (despite non-convex).
Lasso illustration
Evolution of 10 coefficients via ISTA and gradient descent of $f$.

ISTA Noncvx-Pro

\[
 k + 1 = \text{SoftThresh}(k, \lambda X > (X_k y), \lambda)
\]

Noncvx-Pro:

\[
 k + 1 = k + \lambda r_f(k)
\]
Variational Projection: $f(v) \triangleq \min_u F(u, v)$

$u^*(v) \triangleq \arg\min_u F(u, v)$

Envelope theorem: $\nabla f(v) = \nabla_v F(u^*(v), v)$
**VarPro General Idea**

**Variational Projection:** \( f(v) \triangleq \min_u F(u, v) \)

\[ u^*(v) \triangleq \arg\min_u F(u, v) \]

**Envelope theorem:**
\[ \nabla f(v) = \nabla_v F(u^*(v), v) \]

**Implicit diff. of** \( \nabla_u F(u^*, v) = 0: \)
\[ \partial u^*(v) = -\partial^2_u F(u^*, v)^{-1} \partial_{u,v} F(u^*, v) \]

\[ \partial^2 f(v) = \partial^2_v F(u^*, v) - \partial_{u,v} F(u^*, v) \partial^2_u F(u^*, v)^{-1} \partial_{u,v} F(u^*, v) \]
VarPro General Idea

Variational Projection: $f(v) \triangleq \min_u F(u, v)$  \hspace{1cm} $u^*(v) \triangleq \arg\min_u F(u, v)$

Envelope theorem: $\nabla f(v) = \nabla_v F(u^*(v), v)$

Implicit diff. of $\nabla_u F(u^*, v) = 0$: $\partial u^*(v) = -\partial^2_u F(u^*, v)^{-1} \partial_{u,v} F(u^*, v)$

$\partial^2 f(v) = \partial^2_v F(u^*, v) - \partial^2_{u,v} F(u^*, v) \partial^2_u F(u^*, v)^{-1} \partial^2_{u,v} F(u^*, v)$

\[ \Rightarrow \quad \partial F^{-1} = \left( \begin{array}{cc} \partial^2_u F & \partial^2_{u,v} F \\ \partial^2_{u,v} F & \partial^2_v F \end{array} \right)^{-1} = \left( \begin{array}{cc} \cdots & \cdots \\ \cdots & [\partial^2_v f]^{-1} \end{array} \right) \]

$\partial^2_v f$ is better conditionned than $\partial^2_{u,v} F$
\[
\min_{x \in \mathbb{R}^d} \frac{1}{2\lambda} \|Ax - y\|^2 + \|x\|_1
\]

\[x = u \odot v\]

\[
\min_{u,v} F(u, v) \triangleq \frac{1}{\lambda} \|A(u \odot v) - y\|^2 + \|u\|_2^2 + \|v\|_2^2
\]

\[
f(v) \triangleq \min_u \frac{1}{\lambda} \|A(u \odot v) - y\|^2 + \|u\|_2^2 + \|v\|_2^2
\]

\[
\min_v f(v)
\]
Lasso VarPro

\[
\min_{x \in \mathbb{R}^d} \frac{1}{2\lambda} \|Ax - y\|^2 + \|x\|_1
\]

\[
x = u \odot v
\]

\[
\min_{u, v} F(u, v) \triangleq \frac{1}{\lambda} \|A(u \odot v) - y\|^2 + \|u\|_2^2 + \|v\|_2^2
\]

\[
f(v) \triangleq \min_u \frac{1}{\lambda} \|A(u \odot v) - y\|^2 + \|u\|_2^2 + \|v\|_2^2
\]

**Proposition:** \( f \) is smooth and

\[
u^*(v) \triangleq \arg\min_u F(u, v)
\]

\[
= v \odot A^\top (\text{Adag}(v^2)A^\top + \lambda \text{Id}_n)^{-1}y
\]

Envelope theorem:

\[
\nabla f(v) = \nabla_v F(u^*(v), v)
\]

\[
= \frac{1}{\lambda} u^* \odot A^\top (A(u^* \odot v) - y) + v
\]
**Proposition:** $f$ is smooth and

$$u^*(v) \triangleq \arg\min_u F(u, v)$$

$$= v \odot A^\top (A \text{diag}(v^2) A^\top + \lambda \text{Id}_n)^{-1} y$$

Envelope theorem:

$$\nabla f(v) = \nabla_v F(u^*(v), v)$$

$$= \frac{1}{\lambda} u^* \odot A^\top (A (u^* \odot v) - y) + v$$

**Case $\lambda = 0$:**

$$\nabla f(v) = v \odot (1 - (A^\top a)^2)$$

$$a \overset{\text{def.}}{=} (A \text{diag}(v^2) A^\top)^+ y$$
Mild non-convexity

Property 2: "mildly nonconvex"

**Definition**

$v$ is a stationary point if $\nabla f(v) = 0$. It is a strict saddle point if $\nabla f(v) = 0$ but $\lambda_{\text{min}}(\partial^2 f(v)) \neq 0$ does not hold.

**Fact**: Gradient descent always avoids strict saddle points.

For our $f$, all stationary points are either global minimums or strict saddle points (at least one negative eigenvalue).

$\dagger$ Jason D Lee et al. "First-order methods almost always avoid saddle points". In: arXiv preprint arXiv:1710.07406 (2017), Chi Jin et al. "How to escape saddle points efficiently". In: International Conference on Machine Learning. PMLR. 2017, pp. 1724–1732.

$v$ strict saddle point:

$\nabla f(v) = 0$ and $\lambda_{\text{min}}(\partial^2 f(v)) < 0$

→ gradient descent avoids strict saddles.

*Example*: 0 is a strict saddle for Lasso VarPro.
**Mild non-convexity**

\[ v \text{ strict saddle point: } \nabla f(v) = 0 \text{ and } \lambda_{\min}(\partial^2 f(v)) < 0 \]

→ gradient descent avoids strict saddles.

*Example:* 0 is a strict saddle for Lasso VarPro.

**Primal variable:** \[ x \triangleq u^*(v) \odot v \]

**Dual certificate:** \[ \xi \triangleq \frac{1}{\lambda} A^\top (Ax - y) \]

\[ x \text{ global minimum } \iff \xi_I = \text{sign}(x_I) \text{ and } \|\xi_{I^c}\|_\infty \leq 1 \quad I \triangleq \text{supp}(x) \]

**Theorem:** all stationary points are global minima or strict saddles.

\[ \|\partial^2 f(v)\| \leq 1 + 3\|y\|^2 \|A\|^2 / \lambda \]

If \( \nabla f(v) = 0 \),

\[ \text{eig}(\partial^2 f(v)) \in [1 - |\xi_{I^c}|_\infty, 4] \]
**Mild non-convexity**

A strict saddle point:
\[ \nabla f(v) = 0 \text{ and } \lambda_{\text{min}}(\partial^2 f(v)) < 0 \]

→ gradient descent avoids strict saddles.

*Example:* 0 is a strict saddle for Lasso VarPro.

Primal variable:
\[ x \triangleq u^*(v) \odot v \]

Dual certificate:
\[ \xi \triangleq \frac{1}{\lambda} A^\top (Ax - y) \]

\[ x \text{ global minimum } \iff \xi_I = \text{sign}(x_I) \text{ and } \|\xi_{I^c}\|_\infty \leq 1 \quad I \triangleq \text{supp}(x) \]

**Theorem:** all stationary points are global minima or strict saddles.

\[ \|\partial^2 f(v)\| \leq 1 + 3\|y\|^2\|A\|^2/\lambda \]

If \( \nabla f(v) = 0 \),
\[ \text{eig}(\partial^2 f(v)) \in [1 - |\xi_{I^c}|_\infty, 4] \]

"non-degenerate" global minimum:
\[ \xi_I = \text{sign}(x) \text{ and } \|\xi_{I^c}\|_\infty < 1 \implies \text{eig}(\partial^2 f(v)) > 0 \]
General Quadratic Variational Formula

Non-convex regularization:

\[ |x|^\beta \propto \min_{x=uv} u^2 + v^\alpha \quad \beta = \frac{2\alpha}{\alpha + 2} \]

\begin{align*}
\text{convex regul.} & \quad \beta \\
\text{smooth VarPro} & \quad \alpha
\end{align*}
Non-convex regularization:

\[ |x|^\beta \propto \min_{x=uv} u^2 + v^\alpha \quad \beta = \frac{2\alpha}{\alpha + 2} \]

Group Lasso: Disjoint: \( \{1, \ldots, n\} = \bigcup g \ g \)

\[
\|x\|_{1,2} \triangleq \sum_g \|x_g\| = \min_{u_g, v_g = x_g} \sum_g \frac{\|u_g\|^2}{2} + \frac{v_g^2}{2} \quad u_g \in \mathbb{R}^{|g|}, v_g \in \mathbb{R}
\]
General Quadratic Variational Formula

Non-convex regularization:

\[ |x|^{\beta} \propto \min_{x=uv} u^2 + v^\alpha \quad \beta = \frac{2\alpha}{\alpha + 2} \]

Group Lasso: Disjoint: \( \{1, \ldots, n\} = \bigcup_g g \)

\[ \|x\|_{1,2} \triangleq \sum_g \|x_g\| = \min_{u_g, v_g = x_g} \sum_g \frac{\|u_g\|^2}{2} + \frac{v_g^2}{2} \quad u_g \in \mathbb{R}^{|g|}, \ v_g \in \mathbb{R} \]

Trace norm:

\[ \|X\|_* \triangleq \sum_k \sigma_k(X) \quad X \in \mathbb{R}^{r \times s} \]

\[ \|X\|_* = \min_{X=UV} \frac{\|U\|_F^2 + \|V\|_F^2}{2} \quad U \in \mathbb{R}^{r \times \min(r,s)}, \ V \in \mathbb{R}^{\min(r,s) \times s} \]
General Quadratic Variational Formula

[Geman and Reynolds 1992] [Black and Rangarajan 1992]

Non-convex regularization:

\[ |x|^\beta \propto \min_{x=uv} u^2 + v^\alpha \quad \beta = \frac{2\alpha}{\alpha + 2} \]

Group Lasso: Disjoint: \( \{1, \ldots, n\} = \bigcup_g g \)

\[ \|x\|_{1,2} \triangleq \sum_g \|x_g\| = \min_{u_g, v_g = x_g} \sum_g \frac{\|u_g\|^2}{2} + \frac{v_g^2}{2} \]

\( u_g \in \mathbb{R}^{|g|}, v_g \in \mathbb{R} \)

Trace norm:

\[ \|X\|_* \triangleq \sum_k \sigma_k(X) \quad X \in \mathbb{R}^{r \times s} \]

\[ \|X\|_* = \min_{X=UV} \frac{\|U\|_F^2 + \|V\|_F^2}{2} \]

\( U \in \mathbb{R}^{r \times \min(r,s)}, V \in \mathbb{R}^{\min(r,s) \times s} \)

Theorem: \( R(x) = \varphi(x^2) \) with \( \varphi \) concave \iff \( R(x) = \min_{x=uv} \frac{\|u\|^2 + h(v^2)}{2} \)

\( h(\eta) \triangleq 2(-\varphi)^*(1/\eta) \)
Numerics for Lasso

Leukemia (38, 7129)

Earthmover distance on graphs

Quadratic Variational Forms

Bilevel VarPro For Lasso

Continuous Flows

Numerical Results
From Gradient to Quasi-Newton

\[ f(v_k) - \min f \]

\[ f(v) = \min_v F(u, v) \]

Deconvolution:

\[ F(u, v) \]

\[ f(v) = \min_v F(u, v) \]
From Gradient to Quasi-Newton

Gradient descent:

\[ f(v_k) = \min_v F(u, v) \]

\[ f(v) = \min_v F(u, v) \]

\( f \) is smooth \( \rightarrow \) use quasi-Newton

\[ v_{k+1} = v_k - \tau_k B_k \nabla f(v_k) \]

\[ B_k \approx [\partial^2 f(v_k)]^{-1} \]

Secant condition:

\[ \nabla f(v_{k+1}) - \nabla f(v_k) = \tau \partial^2 f(v_k)(v_{k+1} - v_k) + o(\|v_{k+1} - v_k\|) \]

\[ B_{k+1}(\nabla f(v_{k+1}) - \nabla f(v_k)) = v_{k+1} - v_k \]

Rank-1 update:

\[ B_0 = \lambda \text{Id} \quad B_{k+1} = B_k + \sigma hh^\top \]

Near non-degenerate solution: linear convergence.
Lasso

\[\ell \in (n, d) = (22696, 122)\]

(UCI/Adult, income prediction)
Lasso

Leukemia \((n, d) = (38, 7129)\)

[Golub et al. 1999]  Molecular classification of cancer
Group Lasso - MEG+EEG

- 2294 sources locations
  \[
  d = 2294 \times 181
  \]
  Group in time, \(|g| = 181\).
- 301 MEG + 59 EEG sensors
  \[
  n = 360 \times 181
  \]

\[\lambda = \frac{1}{2} \lambda_{\text{max}}\]
\[\lambda = \frac{1}{10} \lambda_{\text{max}}\]
\[\lambda = \frac{1}{100} \lambda_{\text{max}}\]

* Eugene Ndiaye et al. "Gap safe screening rules for sparse multi-task and multi-class models". In: arXiv preprint arXiv:1506.03736 (2015).
Multi-task Learning

\[
\min_{X=(x_t)_{t=1}^T} \frac{1}{2\lambda} \sum_{t=1}^T \|A_t x_t - y_t\|^2 + \|X\|_*
\]

Numerical experiments on 3 datasets

- **Schools**
  - \((T, n, d) = (139, 15362, 27)\)

- **Parkinson**
  - \((T, n, d) = (42, 5875, 19)\)

- **SARCOS**
  - \((T, n, d) = (7, 48933, 21)\)

IRLS-d corresponds to IRLS with \(\varepsilon = 10^{-d}\)
W1 Optimal Transport

\[ W_1(\alpha, \beta) = \min_{\overrightarrow{w}} \left\{ \int \| \overrightarrow{w}(x) \| ; \; \text{div}(\overrightarrow{w}) = \alpha - \beta \right\} \]

Graph:
- Lasso \((\lambda = 0)\)

Discretization
- \(A = \text{div} \in \mathbb{R}^{n \times d}\)
- \(n = |\text{vertex}|\)
- \(d = |\text{edges}|\)

Finite elements:
- Group Lasso \((\lambda = 0)\)
Non-convex Regularization

\[
\min_{Ax=y} \sum_i |x_i|^\beta
\]

VarPro is smooth for \( \beta > 2/3 \)

\[ \beta = 0.7 \quad \beta = 0.8 \quad \beta = 1 \]

Iteration of GD

Compressed sensing

\[ A_{i,j} \sim \mathcal{N}(0, 1) \]

\[ y = Ax_0 \]

\[ \|x_0\|_0 = 25 \]
Mirror Flows and Over-Parameterization

**Entropy function:** $\varphi$ smooth and strongly convex.

\[
x_{k+1} \triangleq (\varphi')^{-1}[\varphi'(x_k) - \tau_k f'(x_k)]
\]

Mirror descent: \[ \tau \to 0 \]

Hessian manifold flow:

\[
\dot{x}(t) = -[\varphi''(x(t))]^{-1} f'(x(t))
\]
The inequality also holds for bounded by Lemma 3.2 norm increases.

\[ \text{Mirror descent:} \quad x_{k+1} \triangleq (\varphi')^{-1}[\varphi'(x_k) - \tau_k f'(x_k)] \]

\[ \text{Hessian manifold flow:} \quad \dot{x}(t) = -[\varphi''(x(t))]^{-1} f'(x(t)) \]

\[ \tau \to 0 \quad t = k\tau \]

**Entropy function:** \( \varphi \) smooth and strongly convex.

Hyperbolic flow on \( \mathbb{R} \):

\[ F(u, v) \triangleq f(uv) \]

\[ (\dot{u}, \dot{v}) = -F'(u, v) \]

\[ \dot{x} = -[\varphi''_\beta(x)]^{-1} f'(x) \]

\[ \varphi_\beta(x) \triangleq x \text{ArcSinh}(\frac{x}{\beta}) - \sqrt{x^2 + \beta^2} + \beta \]

\[ \beta \triangleq |u(0)^2 - v(0)^2|/2 \]
We consider the following assumptions on the divergence-generating function constraint in the methods in the next section.

**Entropy function:** $\varphi$ smooth and strongly convex.

**Mirror descent:**

$$x_{k+1} \triangleq (\varphi')^{-1}[\varphi'(x_k) - \tau_k f'(x_k)]$$

**Hessian manifold flow:**

$$\dot{x}(t) = -[\varphi''(x(t))]^{-1} f'(x(t))$$

**Mirror Flows and Over-Parameterization**

**Hyperbolic flow on $\mathbb{R}$:** $F(u, v) \triangleq f(uv)$

$$(\dot{u}, \dot{v}) = -F'(u, v)$$

$${\dot{x}} = -[\varphi''(x)]^{-1} f'(x)$$

$$\varphi_{\beta}(x) \triangleq x \text{ArcSinh}(\frac{x}{\beta}) - \sqrt{x^2 + \beta^2} + \beta$$

$${\beta} \triangleq |u(0)^2 - v(0)^2|/2$$

**Fisher-Rao flow on $\mathbb{R}^+$:** $F(u) \triangleq f(u^2)$

$$\dot{u} = -F'(u)$$

$${\dot{x}} = -[\varphi''(x)]^{-1} f'(x)$$

$$\varphi_{+}(x) \triangleq x \log(x) - x + 1$$
Hadamard Flow

\[
\min_x f(x) \triangleq \frac{1}{2\lambda} \|Ax - y\|^2 + \|x\|_1
\]

\[x = u \odot v\]

\[
\min_{u,v} F(u,v) \triangleq \frac{1}{\lambda} \|A(u \odot v) - y\|^2 + \|u\|_2^2 + \|v\|_2^2
\]

\[
\min_x f_\beta(x) \triangleq \frac{1}{2\lambda} \|Ax - y\|^2 + \sum_i \sqrt{\beta^2 + x_i^2}
\]
Hadamard Flow

\[
\min_x f(x) \triangleq \frac{1}{2\lambda} \|Ax - y\|^2 + \|x\|_1
\]

\[
x = u \odot v
\]

\[
\min_{u,v} F(u, v) \triangleq \frac{1}{\lambda} \|A(u \odot v) - y\|^2 + \|u\|^2 + \|v\|^2
\]

\[
\min_x f_\beta(x) \triangleq \frac{1}{2\lambda} \|Ax - y\|^2 + \sum_i \sqrt{\beta^2 + x_i^2}
\]

**Theorem:**

\[
\dot{x} = -[\varphi''_{\beta_t}(x)]^{-1} \nabla f_{\beta_t}(x)
\]

\[
x = u \odot v
\]

\[
(\dot{u}, \dot{v}) = -\nabla F(u, v)
\]

\[
\beta_t \triangleq |u(0)^2 - v(0)^2| e^{-2\lambda t}
\]

\[
\varphi_\beta(x) \triangleq x \text{ArcSinh}(\frac{x}{\beta}) - \sqrt{x^2 + \beta^2 + \beta}
\]
Hadamard Flow

\[
\min_x f(x) \triangleq \frac{1}{2\lambda} \|Ax - y\|^2 + \|x\|_1 \quad \Rightarrow \quad \min_{u,v} F(u, v) \triangleq \frac{1}{\lambda} \|A(u \odot v) - y\|^2 + \|u\|_2^2 + \|v\|_2^2
\]

\[
\min_x f_\beta(x) \triangleq \frac{1}{2\lambda} \|Ax - y\|^2 + \sum_i \sqrt{\beta^2 + x_i^2}
\]

**Theorem:**

\[
\dot{x} = -[\varphi''(x)]^{-1} \nabla f_{\beta_t}(x) \quad \Rightarrow \quad \dot{x} = u \odot v \quad (\dot{u}, \dot{v}) = -\nabla F(u, v)
\]

\[
\beta_t \triangleq |u(0)^2 - v(0)^2|e^{-2\lambda t}
\]

Discrete flow:

\[
x_k = u_k \odot v_k \quad (u_{k+1}, v_{k+1}) \triangleq (u_{k+1}, v_{k+1}) - \tau \nabla F(u_k, v_k)
\]

**Theorem:**

\[
\min_{\ell \leq k} \|\nabla F(u_\ell, v_\ell)\| = O(1/\sqrt{k})
\]

\[
f(x_k) - \lim_{\ell \to +\infty} f(x_\ell) \leq C \sum_{\ell > k} \|\nabla F(u_\ell, v_\ell)\|^2
\]

Open problem: mirror-like interpretation for VarPro?
Conclusion

Practice

- Trade smoothness / non-convexity
- Simple algorithm
- Handles small and even 0 value of $\lambda$

? Efficient sparse linear system?
? Including pruning?
Conclusion

Practice

- Trade smoothness / non-convexity
- Simple algorithm
- Handles small and even 0 value of $\lambda$
- Efficient sparse linear system?
- Including pruning?

Theory

- Operates a reconditioning
- Mild non-convexity
- Hyperbolic geometry is better than Euclidean
- Fine-grid analysis?
- Mirror-flow analysis?