

# Steepest Descent Analysis for Unregularized Linear Prediction with Strictly Convex Penalties

Matus Telgarsky <[mtelgars@cs.ucsd.edu](mailto:mtelgars@cs.ucsd.edu)>

# Setup.

- ▶ Primal problem

$$\inf \{f(A\lambda) : \lambda \in \mathbb{R}^n\}$$

with matrix  $A \in \mathbb{R}^{m \times n}$  and function  $f : \mathbb{R}^m \rightarrow \mathbb{R}$  satisfying:

- ▶  $f$  strictly convex, twice continuously differentiable, bounded below, finite everywhere,
- ▶  $f \circ A$  has Lipschitz gradients.
- ▶ Steepest descent: start with  $\lambda_0 := \mathbf{0}_n$ , then repeatedly
  - ▶ choose steepest direction  $v_t$  wrt some norm  $\|\cdot\|$ :

$$v_t \in \underset{v}{\text{Arg max}} \{ \langle v, \nabla(f \circ A)(\lambda_{t-1}) \rangle : \|v\| \leq 1 \};$$

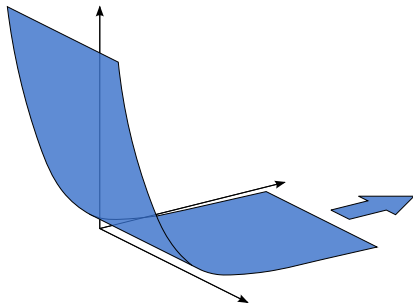
- ▶ and line search for  $\lambda_t := \lambda_{t-1} + \alpha_t v_t$ .

# Difficulties.

- ▶ Primal problem

$$\inf_{\lambda \in \mathbb{R}^n} f(A\lambda) = \inf_{\lambda \in \mathbb{R}^n} (f \circ A)(\lambda).$$

- ▶ Let's prove some convergence rates. However:
  - ▶  $f \circ A$  not necessarily strictly convex,
  - ▶  $f \circ A$  may fail to have a minimizer!
    - ▶ Problematic for other methods as well.



# Outline.

- ▶ Setup.
- ▶ A template theorem.
- ▶ Specific  $(f, A)$ .

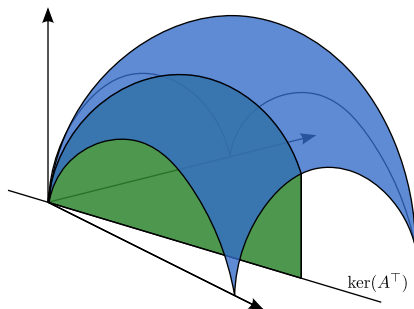
# Dual problem.

- ▶ Duality

$$\inf \{f(x) : x \in \text{im}(A)\} = \sup \left\{ -f^*(\phi) : \phi \in \ker(A^\top) \right\}.$$

Crucially,

- ▶ troublesome  $A$  is out of the objective, and
- ▶ there is always a (unique) maximizer.



## Dual potential function.

- ▶ Consider potential function / dual distance

$$\Psi(\phi_t) := \Psi(\nabla f(A\lambda_t)) := \inf \left\{ \|\phi_t - \psi\|_* : \psi \in S \cap \ker(A^\top) \right\},$$

where  $\|\cdot\|$  and  $S \subseteq \mathbb{R}^m$  chosen for technical convenience.

## Upper bounding dual distance.

**(Potential function upper bound.)** Magic quantity is

$$\gamma := \inf \left\{ \frac{\overbrace{\|A^\top(\phi - \psi)\|_*}_{\|A^\top \phi\|_*}, \psi \in \ker(A^\top)}{\underbrace{\inf\{\|\phi - \psi\|_* : \psi \in S \cap \ker(A^\top)\}}_{\Psi(\phi)}} : \phi \in S \setminus \ker(A^\top) \right\}.$$

Using line search guarantee, Lipschitz gradient constant  $L_t$  over current sublevel set, and some conditions on  $S$ :  $\gamma > 0$  and

$$\Psi(\phi_t)^2 \leq \frac{1}{\gamma^2} \|\nabla(f \circ A)(\lambda_t)\|^2 \leq \frac{6L_t}{\gamma^2} (f(A\lambda_t) - f(A\lambda_{t+1})).$$

## Template theorem.

**(Assumed potential function lower bound.)** Let  $D_h$  be the Bregman divergence wrt  $h$ . Suppose exist  $C > 0$  and  $k \in \{1, 2\}$  so that, for all  $t$ ,

$$CL_t \left( \inf_{\psi \in S \cap \ker(A^\top)} D_{f^*}(\psi, \phi_t) \right)^k \leq \underbrace{\left( \inf_{\psi \in S \cap \ker(A^\top)} \|\phi_t - \psi\|_* \right)^2}_{\Psi(\phi_t)}.$$

Then, for any  $\epsilon > 0$ :

- ▶ if  $k = 1$ , then  $\mathcal{O}(\ln(\frac{1}{\epsilon}))$  iterations suffice.
- ▶ if  $k = 2$ , then  $\mathcal{O}(\frac{1}{\epsilon})$  iterations suffice;



# Outline.

- ▶ Setup.
- ▶ A template theorem.
- ▶ Specific  $(f, A)$ .

# Minimizers.

- $f \circ A$  has minimizers,  $\nabla^2 f \succ \mathbf{0}_{m \times m}$ .
- $\implies f + \iota_{\text{im}(A)}$  has compact level sets.
- $\implies f + \iota_{\text{im}(A)}$  strongly convex over every sublevel set.
- $\implies$ exists  $c > 0$  with  $D_{f^*}(\psi, \phi_t) \leq \frac{1}{c} \|\phi_t - \psi\|_*^2$ .
- $\implies$ template theorem satisfied with  $k = 1$ .
- $\implies$ rate  $\mathcal{O}(\ln(1/\epsilon))$ .

**Examples:** Linear regression problems

$$\sum_{i=1}^m (y_i - \langle x_i, \lambda \rangle)^2 \quad \text{and} \quad \sum_{i=1}^m \ln(\cosh((y_i - \langle x_i, \lambda \rangle))),$$

boosting when minimizer exists, quadratics with positive *semi*-definite Hessians.

# Unattainability in Boosting I.

- ▶ For boosting's  $f$ , any sequence within  $\mathbb{R}_{--}^m$  with unbounded growth in each coordinate has a minimizing subsequence.
- ▶ Suppose that  $\text{im}(A) \cap \mathbb{R}_{--}^m \neq \emptyset$ , and  $f$  satisfies a **flattening condition**: for any sublevel set  $S_\epsilon$ , there exist  $\eta > 0$  and  $\beta > 0$  so that for any  $x \in S_\epsilon$ ,

$$\eta^{-1}L_t \leq f(x) - \inf_z f(z) \leq \beta \|\nabla f(x)\|_1.$$

Then the template theorem is satisfied with  $k = 1$ , giving rate  $\mathcal{O}(\ln(\frac{1}{\epsilon}))$ .

- ▶ In the boosting literature,  $\text{im}(A) \cap \mathbb{R}_{--}^m \neq \emptyset$  is *the (sample-specific) weak learning assumption*.

## Unattainability in Boosting II.

- ▶ For a general choice of  $A$ , every convergent minimizing sequence  $(\lambda_i)_{i=1}^{\infty}$  will necessarily have  $(A\lambda_i) \uparrow \infty$  for a subset of  $[m]$ , the remaining converging to finite points.
- ▶ Can break objective in two, execute previous analysis on each piece, and stitch back together to achieve  $k = 2$ , thus rate  $\mathcal{O}(\frac{1}{\epsilon})$ .

## Summary.

- ▶ Linear convergence when  $f \circ A$  has minimizers,  $f$  satisfies (..).
- ▶ A loose strategy to control convergence w/o minimizers.
- ▶ What is the right flattening condition?

# Summary.

- ▶ Linear convergence when  $f \circ A$  has minimizers,  $f$  satisfies (..).
- ▶ A loose strategy to control convergence w/o minimizers.
- ▶ What is the right flattening condition?

Thanks!!

# Summary.

- ▶ Linear convergence when  $f \circ A$  has minimizers,  $f$  satisfies (..).
- ▶ A loose strategy to control convergence w/o minimizers.
- ▶ What is the right flattening condition?

Thanks!!