Optimization for Data Science

Lec 06: Mirror Descent

Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

Problem

Constrained minimization:

$$f_\star = \inf_{\mathbf{w} \in C \subseteq \mathsf{V}} f(\mathbf{w})$$

- f: convex and possibly nonsmooth
- C: convex constraint
- V: vector space that $\mathbf{w}$ lives in, e.g. $\mathbb{R}^d$ with Euclidean norm $\|\cdot\|_2$
- When f' is L-Lipschitz, (projected) gradient descent yields $\frac{\mathsf{L}\|\mathbf{w}_0 - \mathbf{w}\|_2^2}{2t}$
- When f is L-Lipschitz, (projected) subgradient yields $\frac{\mathsf{L}\|\mathbf{w}_0 - \mathbf{w}\|_2}{\sqrt{t}}$

How We Measure Things Matters

$$\min_{\mathbf{w}\in\Delta} \sum_{j=1}^{d} f_j(w_j),$$

- Each univariate function $f_j : \mathbb{R} \to \mathbb{R}$ is 1-Lipschitz continuous.
- The sum $f : \mathbb{R}^d \to \mathbb{R}$ is $\sqrt{d}$-Lipschitz continuous w.r.t. the Euclidean norm:

$$\|f'\|_2^2 = \sum_j (f_j')^2 \leq d.$$

- The diameter $\|\mathbf{w}_0 - \mathbf{w}\|_2 \leq \sqrt{2}$.
- Applying subgradient we obtain a convergence rate of $\sqrt{\frac{2d}{t}}$
- But, we also have $\|f'\|_\infty = \max_j |f_j'| \leq 1$
- The diameter $\|\mathbf{w}_0 - \mathbf{w}\|_1 = \|\mathbf{w}_0 - \mathbf{w}\|_\infty^\circ \leq 2$
- Possible to achieve the convergence rate $\frac{2}{\sqrt{t}}$ by changing the norm?

What Makes Incremental Update Possible?

So far, all of our updates are of the following (additive) incremental form:

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \cdot \mathbf{g},$$

which is so natural that we often forget what makes it even mathematically possible:

- The scalar multiplication of the step size η to $\mathbf{g}$
- The negation $-$
- And the addition of $\mathbf{w}$ with $-\eta \cdot \mathbf{g}$

$\boxed{\text{These operations are possible because } \mathbf{w} \text{ and } \mathbf{g} \text{ are from the same vector space}}$

- From now on $f'(\mathbf{w})$ lives in a dual space V^*
- Need a way to pull things back and forth: $\mathsf{J} : \mathsf{V} \to \mathsf{V}^*, \quad \mathsf{J}^{-1} : \mathsf{V}^* \to \mathsf{V}$
- With the Euclidean norm $\|\cdot\|_2$, we may simply take $\mathsf{J} = \mathsf{J}^* = \mathrm{Id}$

Algorithm 1: Winnow

Input: $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n] \in \mathbb{R}^{\mathsf{p} \times \mathsf{n}}$, threshold $\delta \geq 0$, step size $\eta > 0$, initialize $\mathbf{w} \in \operatorname{int}\Delta_{\mathsf{p}-1}$

Output: approximate solution $\mathbf{w}$

```
for t = 1, 2, ... do
    receive training example index I_t ∈ {1, ..., n}      // index I_t can be random
    if ⟨a_{I_t}, w⟩ ≤ δ then
        w ← w ⊙ exp(η a_{I_t})                            // update only when making a mistake
        w ← w / ‖w‖_1                                     // normalize
```

$$\ln \mathbf{w} \leftarrow \ln \mathbf{w} + \eta \cdot \mathbf{a}_{I_t}, \quad \text{where} \quad \mathsf{J}(\mathbf{w}) = \ln(\mathbf{w})$$

N. Littlestone. "Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm". *Machine Learning*, vol. 2 (1988), pp. 285–318.

Online Prediction

- n experts, each of whom provides a prediction x_i, collectively as $\mathbf{x} \in \mathbb{R}^n$
- Form our own opinion by averaging $\hat{y} = \langle \mathbf{w}, \mathbf{x} \rangle$, $\mathbf{w} \in \Delta$
- Suffer a loss, say the square loss $\ell(\mathbf{w}; \mathbf{x}, y) = (y - \hat{y})^2$
- Repeat the game for $t = 1, \ldots, T$ rounds

$$\text{Regret} := \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2 - \min_{\mathbf{w}\in\Delta}\frac{1}{T}\sum_{t=1}^{T}(y_t - \langle \mathbf{w}, \mathbf{x}_t \rangle)^2, \quad \text{where} \quad \hat{y}_t = \langle \mathbf{w}_t, \mathbf{x}_t \rangle .$$

Exponentiated Gradient (EG)

$$\tilde{\mathbf{w}}_{t+1} = \mathbf{w}_t \odot \exp(-\eta_t \cdot \ell'(\hat{y}_t - y_t)\mathbf{x}_t)$$
$$\mathbf{w}_{t+1} = \frac{\tilde{\mathbf{w}}_{t+1}}{\langle \mathbf{1}, \tilde{\mathbf{w}}_{t+1} \rangle}$$

- Diminishing regret on the order of $O(\sqrt{\frac{\ln n}{T}})$, assuming $\|\mathbf{x}_t\|_\infty \leq 1$ and $y_t \in [0, 1]$
- No assumption on how the sequence $(\mathbf{x}_t, y_t)$ is generated; can even be adversarial
- Setting $\mathbf{w} = \mathbf{e}_i$: EG performs no worse than the best expert in hindsight for big T
- Can consult a large number of experts: dependence on n is only logarithmic
- Gradient descent achieves $O(\frac{1}{\sqrt{T}})$ under the assumption $\|\mathbf{x}_t\|_2 \leq 1$

J. Kivinen and M. K. Warmuth. "Exponentiated Gradient versus Gradient Descent for Linear Predictors". *Information and Computation*, vol. 132, no. 1 (1997), pp. 1–63.

Two Choices

- We have a mismatch between $\mathbf{w} \in \mathsf{V}$ and $f'(\mathbf{w}) \in \mathsf{V}^*$
- We use a duality (mirror) map $\mathsf{J} : \mathsf{V} \to \mathsf{V}^*, \quad \mathsf{J}^{-1} : \mathsf{V}^* \to \mathsf{V}$

1. Update in the gradient space V^* and pull the update back to the input space V:

$$\mathbf{w}_{t+1} = \mathsf{J}^{-1}[\mathsf{J}(\mathbf{w}_t) - \eta_t \cdot f'(\mathbf{w}_t)]$$
$$\mathbf{w}^*_{t+1} = \mathbf{w}^*_t - \eta_t \cdot f'(\mathsf{J}^{-1}\mathbf{w}^*_t), \quad \text{where} \quad \mathbf{w}^*_t := \mathsf{J}(\mathbf{w}_t), \; \mathbf{w}_t = \mathsf{J}^{-1}(\mathbf{w}^*_t)$$

2. Pull the gradient back to the input space V and do the update there directly:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \cdot \mathsf{J}^{-1}(f'(\mathbf{w}_t)).$$

Legendre function

We call a continuous convex function h Legendre if

- Its domain has nonempty interior, i.e., $\mathrm{int}(\mathrm{dom}\, h) \neq \emptyset$
- h is differentiable on $\mathrm{int}(\mathrm{dom}\, h)$
- $\|h'(\mathbf{w})\| \to \infty$ as $\mathbf{w} \to \partial\, \mathrm{dom}\, h$
- h is strictly convex on $\mathrm{int}(\mathrm{dom}\, h)$

Theorem: J $= h'$

h' is a topological isomorphism, i.e. it is continuous and its inverse is also continuous.

Below, we will choose a norm $\|\cdot\|$ and a Legendre function h that is 1-strongly convex w.r.t. $\|\cdot\|$, i.e.

$$\mathrm{D}_h(\mathbf{w}, \mathbf{z}) := h(\mathbf{w}) - h(\mathbf{z}) - \langle \mathbf{w} - \mathbf{z}; \nabla h(\mathbf{z}) \rangle \geq \tfrac{1}{2}\|\mathbf{w} - \mathbf{z}\|^2.$$

Example: (Squared) Euclidean distance

Let $h(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$. Then, h is Legendre and its induced Bregman divergence $\mathsf{D}_h(\mathbf{w}, \mathbf{z}) = \frac{1}{2}\|\mathbf{w} - \mathbf{z}\|_2^2$ is the (square) Euclidean distance. We have $\mathsf{J}(\mathbf{w}) = h'(\mathbf{w}) = \mathbf{w}$ and of course $\mathsf{J}^{-1} = \mathsf{J}$.

Example: KL and Pinsker

Consider the KL function $h(\mathbf{w}) = \sum_j w_j \ln w_j - w_j$, where $0 \ln 0 := 0$. It is Legendre and its induced Bregman divergence D_h is known as the KL divergence:

$$\forall \mathbf{w}, \mathbf{z} \geq \mathbf{0}, \quad \mathsf{KL}(\mathbf{w}, \mathbf{z}) = \sum_j w_j \ln \frac{w_j}{z_j} - w_j + z_j,$$

which is 1-strongly convex w.r.t. the ℓ_1 norm (restricted to the simplex):

$$\forall \mathbf{w}, \mathbf{z} \in \Delta, \quad \mathsf{KL}(\mathbf{w}, \mathbf{z}) \geq \tfrac{1}{2}\|\mathbf{w} - \mathbf{z}\|_1^2,$$

also known as Pinsker's inequality in information theory.

Algorithm 2: Mirror descent

Input: $\mathbf{w}_0 \in C$, Legendre function h

for $t = 0, 1, \ldots$ **do**

  compute (sub)gradient $f'(\mathbf{w}_t)$

  choose step size $\eta_t > 0$

  $h'(\mathbf{z}_{t+1}) = h'(\mathbf{w}_t) - \eta_t \cdot f'(\mathbf{w}_t)$ // update in the gradient space

  $\mathbf{w}_{t+1} \leftarrow \operatorname*{argmin}_{\mathbf{w}\in C} \mathsf{D}_h(\mathbf{w}, \mathbf{z}_{t+1})$ // projecting back to the constraint

Key insight (note the similarity as before):

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in C} \; f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t; f'(\mathbf{w}_t)\rangle + \frac{1}{\eta_t}\mathsf{D}_h(\mathbf{w}, \mathbf{w}_t)$$

$$\geq \; f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t; f'(\mathbf{w}_t)\rangle + \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|^2$$

$$= \operatorname*{argmin}_{\mathbf{w}\in C} \; \mathsf{D}_h(\mathbf{w}, \mathbf{z}_{t+1}), \quad \text{where} \quad h'(\mathbf{z}_{t+1}) = h'(\mathbf{w}_t) - \eta_t \cdot f'(\mathbf{w}_t),$$

A. Nemirovski and D. B. Yudin. "Efficient methods for solving large-scale convex programming problems". *Ekonomika i matematicheskie metody*, vol. 15, no. 1 (1979), pp. 133–152, A. Beck and M. Teboulle. "Mirror descent and nonlinear projected subgradient methods for convex optimization". *Operations Research Letters*, vol. 31, no. 3 (2003), pp. 167–175.

EG $\in$ MD

- Let $C = \Delta$ and h be KL
- We compute the Bregman projection:

$$
\begin{aligned}
\operatorname*{argmin}_{\mathbf{w}\in\Delta} \ \mathsf{KL}(\mathbf{w}, \mathbf{z}) &= \sum_j w_j \log \frac{w_j}{z_j} - w_j + z_j \\
&= \sum_j w_j \log \frac{w_j}{z_j / \langle \mathbf{1}, \mathbf{z} \rangle} - \log \langle \mathbf{1}, \mathbf{z} \rangle - 1 + \langle \mathbf{1}, \mathbf{z} \rangle \\
&\equiv \mathsf{KL}(\mathbf{w}, \tfrac{\mathbf{z}}{\langle \mathbf{1}, \mathbf{z} \rangle})
\end{aligned}
$$

- $h'(\mathbf{w}) = \ln \mathbf{w}$ while $(h')^{-1}(\mathbf{g}) = \exp(\mathbf{g})$, all component-wise
- The mirror descent step reduces to:

$$
\mathbf{z}_{t+1} = (h')^{-1}(h'(\mathbf{w}_t) - \eta_t \cdot f'(\mathbf{w}_t)) = \mathbf{w}_t \odot \exp(-\eta_t f'(\mathbf{w}_t)), \quad \mathbf{w}_{t+1} = \tfrac{\mathbf{z}_{t+1}}{\langle \mathbf{1}, \mathbf{z}_{t+1} \rangle}
$$

choose a Legendre function h that matches the "geometry" (i.e. norm) of the constraint set C, so that projection is trivial

Theorem: convergence of mirror descent for smooth function

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and L-smooth (w.r.t. some norm $\|\cdot\|$), $C \subseteq \mathbb{R}^d$ be closed convex, and η_t is chosen suitably, then for all $\mathbf{w} \in C$ and $t \geq 1$, the mirror descent iterates $\{\mathbf{w}_t\} \subseteq C$ satisfy:

$$f(\mathbf{w}_t) \leq f(\mathbf{w}) + \frac{\mathsf{D}_h(\mathbf{w}, \mathbf{w}_0)}{t\bar{\eta}_t}, \quad \text{where} \quad \bar{\eta}_t := \frac{1}{t}\sum_{s=0}^{t-1} \eta_s,$$

$\mathsf{D}_h(\mathbf{w}, \mathbf{w}_0) \geq \frac{1}{2}\|\mathbf{w} - \mathbf{w}_0\|^2$ for some 1-strongly convex Legendre function h.

- Again, the rate of convergence does not depend on d, the dimension!
- Proof is literally the same as that of projected gradient
- Choosing $\eta_t \equiv 1/\mathsf{L}$ we obtain $f(\mathbf{w}_t) - f(\mathbf{w}) \leq \frac{\mathsf{L}\mathsf{D}_h(\mathbf{w},\mathbf{w}_0)}{t}$
- As before, the dependence on L and $\mathbf{w}_0$ makes intuitive sense.

$$
\begin{aligned}
f(\mathbf{w}_{t+1}) &\le f(\mathbf{w}_t) + \langle \mathbf{w}_{t+1} - \mathbf{w}_t; f'(\mathbf{w}_t)\rangle + \tfrac{1}{\eta_t}\mathsf{D}_h(\mathbf{w}_{t+1}, \mathbf{w}_t) \\
&\le f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t; f'(\mathbf{w}_t)\rangle + \tfrac{1}{\eta_t}\mathsf{D}_h(\mathbf{w}, \mathbf{w}_t) - \tfrac{1}{\eta_t}\mathsf{D}_h(\mathbf{w}, \mathbf{w}_{t+1}) \\
&\le f(\mathbf{w}) + \tfrac{1}{\eta_t}\mathsf{D}_h(\mathbf{w}, \mathbf{w}_t) - \tfrac{1}{\eta_t}\mathsf{D}_h(\mathbf{w}, \mathbf{w}_{t+1}),
\end{aligned}
$$

where the second inequality follows from $\mathbf{w}_{t+1}$ being the Bregman projection to the convex set C, and the last inequality is due to the convexity of f.

Take $\mathbf{w} = \mathbf{w}_t$ we see that

$$
f(\mathbf{w}_{t+1}) \le f(\mathbf{w}_t),
$$

i.e., the algorithm is descending. Summing from $t = 0$ to $t = T - 1$:

$$
T\bar{\eta}_T \cdot [f(\mathbf{w}_T) - f(\mathbf{w})] \le \sum_{t=0}^{T-1} \eta_t [f(\mathbf{w}_{t+1}) - f(\mathbf{w})] \le \mathsf{D}_h(\mathbf{w}, \mathbf{w}_0).
$$

Dividing both sides by $T\bar{\eta}_T$ completes the proof.

Theorem: convergence of mirror descent for nonsmooth function

Let $C \subseteq \mathbb{R}^d$ be closed convex and $f : C \to \mathbb{R}$ be L-Lipschitz continuous convex (w.r.t. some norm $\|\cdot\|$). Start with $\mathbf{w}_0 \in C$, for any $\mathbf{w} \in C$, we have:

$$\min_{0 \le t \le T-1} f(\mathbf{w}_t) - f(\mathbf{w}) \le \sum_{t=0}^{T-1} \frac{\eta_t}{\sum_{s=0}^{T-1} \eta_s} (f(\mathbf{w}_t) - f(\mathbf{w})) \le \frac{2\mathsf{D}_h(\mathbf{w}, \mathbf{w}_0) + \mathsf{L}^2 \sum_{t=0}^{T-1} \eta_t^2}{2 \sum_{s=0}^{T-1} \eta_s},$$

where $\mathsf{D}_h(\mathbf{w}, \mathbf{w}_0) \ge \frac{1}{2}\|\mathbf{w} - \mathbf{w}_0\|^2$ for some 1-strongly convex Legendre function h.

- The bound on the right-hand side vanishes iff $\sum_t \eta_t \to \infty$ and $\eta_t \to 0$
- If we fix a tolerance $\epsilon > 0$ beforehand, then setting $\eta_t = c/\mathsf{L}^2 \cdot \epsilon$ for some constant $c \in]0, 2[$ leads to $\min_{0 \le t \le T-1} f(\mathbf{w}_t) - f(\mathbf{w}) \le \epsilon$, as long as $T \ge \frac{2\mathsf{L}^2 \mathsf{D}_h(\mathbf{w}, \mathbf{w}_0)}{c(2-c)} \cdot \frac{1}{\epsilon^2}$
- The same claim holds for $\bar{\mathbf{w}}_T := \sum_{t=0}^{T-1} \frac{\eta_t}{\sum_{s=0}^{T-1} \eta_s} \mathbf{w}_t$

As in the previous proof, since $\mathbf{w}_{t+1}$ is the Bregman projection, we have

$$\begin{aligned}
\langle \mathbf{w}; f'(\mathbf{w}_t)\rangle + \tfrac{1}{\eta_t}\mathrm{D}_h(\mathbf{w},\mathbf{w}_t) &\geq \langle \mathbf{w}_{t+1}; f'(\mathbf{w}_t)\rangle + \tfrac{1}{\eta_t}\mathrm{D}_h(\mathbf{w}_{t+1},\mathbf{w}_t) + \tfrac{1}{\eta_t}\mathrm{D}_h(\mathbf{w},\mathbf{w}_{t+1})\\
\langle \mathbf{w}-\mathbf{w}_t; f'(\mathbf{w}_t)\rangle + \tfrac{1}{\eta_t}\mathrm{D}_h(\mathbf{w},\mathbf{w}_t) &\geq \langle \mathbf{w}_{t+1}-\mathbf{w}_t; f'(\mathbf{w}_t)\rangle + \tfrac{1}{\eta_t}\mathrm{D}_h(\mathbf{w}_{t+1},\mathbf{w}_t) + \tfrac{1}{\eta_t}\mathrm{D}_h(\mathbf{w},\mathbf{w}_{t+1})\\
f(\mathbf{w}) - f(\mathbf{w}_t) + \tfrac{1}{\eta_t}\mathrm{D}_h(\mathbf{w},\mathbf{w}_t) &\geq -\|\mathbf{w}_{t+1}-\mathbf{w}_t\|\cdot\|f'(\mathbf{w}_t)\|_\circ + \tfrac{1}{2\eta_t}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2 + \tfrac{1}{\eta_t}\mathrm{D}_h(\mathbf{w},\\
f(\mathbf{w}) - f(\mathbf{w}_t) + \tfrac{1}{\eta_t}\mathrm{D}_h(\mathbf{w},\mathbf{w}_t) &\geq \eta_t\|f'(\mathbf{w}_t)\|_\circ^2/2 + \tfrac{1}{\eta_t}\mathrm{D}_h(\mathbf{w},\mathbf{w}_{t+1}).
\end{aligned}$$

Telescoping we obtain

$$\mathrm{D}_h(\mathbf{w},\mathbf{w}_T) \leq \mathrm{D}_h(\mathbf{w},\mathbf{w}_0) + \sum_{t=0}^{T-1}\eta_t^2\|f'(\mathbf{w}_t)\|_\circ^2/2 + \sum_{t=0}^{T-1}\frac{\eta_t}{\sum_{s=0}^{T-1}\eta_s}(f(\mathbf{w}) - f(\mathbf{w}_t))\cdot\sum_{s=0}^{T-1}\eta_s.$$

Thus,

$$\min_{0\leq t\leq T-1} f(\mathbf{w}_t) - f(\mathbf{w}) \leq \sum_{t=0}^{T-1}\frac{\eta_t}{\sum_{s=0}^{T-1}\eta_s}(f(\mathbf{w}_t) - f(\mathbf{w})) \leq \frac{2\mathrm{D}_h(\mathbf{w},\mathbf{w}_0) + \mathrm{L}^2\sum_{t=0}^{T-1}\eta_t^2}{2\sum_{s=0}^{T-1}\eta_s}.$$

Extending to Composite

$$\min_{\mathbf{w}} \; f(\mathbf{w}), \quad \text{where} \quad f(\mathbf{w}) = \ell(\mathbf{w}) + r(\mathbf{w})$$

Algorithm 3: Composite mirror descent

Input: $\mathbf{w}_0$, functions ℓ and r, Legendre function h

for $t = 0, 1, \ldots$ **do**

- compute (sub)gradient $\ell'(\mathbf{w}_t)$ `// can be stochastic`
- choose step size $\eta_t > 0$
- $h'(\mathbf{z}_{t+1}) = h'(\mathbf{w}_t) - \eta_t \cdot \ell'(\mathbf{w}_t)$ `// gradient step w.r.t.` ℓ
- $\mathbf{w}_{t+1} \leftarrow \operatorname*{argmin}_{\mathbf{w}} \frac{1}{\eta_t} \mathsf{D}_h(\mathbf{w}, \mathbf{z}_{t+1}) + r(\mathbf{w})$ `// proximal step w.r.t.` r

J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. "Composite Objective Mirror Descent". In: *Proceedings of the 23rd Annual Conference on Learning Theory*. 2010, J. C. Duchi, A. Agarwal, M. Johansson, and M. I. Jordan. "Ergodic Mirror Descent". *SIAM Journal on Optimization*, vol. 22, no. 4 (2012), pp. 1549–1578.