CS281B/Stat241B. Statistical Learning Theory. Lecture 16.

Peter Bartlett

Recall: Online Prediction

- Repeated game:

$$\text{Decision method plays } a_t \in \mathcal{A}$$
$$\text{World reveals } \ell_t \in \mathcal{L}$$

- Minimax regret is the value of the game:

$$\min_{a_1 \in \mathcal{A}} \max_{\ell_1 \in \mathcal{L}} \cdots \min_{a_n \in \mathcal{A}} \max_{\ell_n \in \mathcal{L}} \left(\hat{L}_n - L_n^* \right).$$

Online Convex Optimization

1. Problem formulation
2. Empirical minimization fails.
3. Gradient algorithm.
4. Regularized minimization
   - Bregman divergence
   - Regularized minimization $\Leftrightarrow$ minimizing latest loss and divergence from previous decision
   - Constrained minimization equivalent to unconstrained plus Bregman projection
   - Linearization
   - Mirror descent
5. Regret bounds

Recall: A Regularization Viewpoint

- Suppose ℓ_t is linear: $\ell_t(a) = g_t \cdot a$, and $\mathcal{A} = \mathbb{R}^d$.
- Then we can view the gradient step

$$a_{t+1} = a_t - \eta \nabla \ell_t(a_t)$$

as minimizing the regularized criterion

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} \left(\eta \sum_{s=1}^{t} \ell_s(a) + \frac{1}{2} \|a\|^2 \right).$$

Recall: Regularization

Regularized minimization

Consider the family of strategies of the form:

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} \left(\eta \sum_{s=1}^{t} \ell_s(a) + R(a) \right).$$

The regularizer $R : \mathbb{R}^d \to \mathbb{R}$ is strictly convex and differentiable.

- R keeps the sequence of a_ts stable: it diminishes ℓ_t's influence.
- We can view the choice of a_{t+1} as trading off two competing forces: making $\ell_t(a_{t+1})$ small, and keeping a_{t+1} close to a_t.

Recall: Regularization

In the unconstrained case ($\mathcal{A} = \mathbb{R}^d$), regularized minimization is equivalent to minimizing the latest loss and the distance to the previous decision. The appropriate notion of distance is the Bregman divergence $D_{\Phi_{t-1}}$:

Define

$$\Phi_0 = R,$$
$$\Phi_t = \Phi_{t-1} + \eta \ell_t,$$

so that

$$\begin{aligned} a_{t+1} &= \arg\min_{a \in \mathcal{A}} \left(\eta \sum_{s=1}^{t} \ell_s(a) + R(a) \right) \\ &= \arg\min_{a \in \mathcal{A}} \Phi_t(a). \end{aligned}$$

Recall: Bregman Divergence

Definition 1. *For a strictly convex, differentiable $\Phi : \mathbb{R}^d \to \mathbb{R}$, the Bregman divergence wrt Φ is defined, for $a, b \in \mathbb{R}^d$, as*

$$D_\Phi(a, b) = \Phi(a) - (\Phi(b) + \nabla\Phi(b) \cdot (a - b)) .$$

$D_\Phi(a, b)$ is the difference between $\Phi(a)$ and the value at a of the linear approximation of Φ about b. (PICTURE)

Example:

- $\Phi(a) = \frac{1}{2}\|a\|^2$: $D_\Phi(a, b) = \frac{1}{2}\|a - b\|^2$.

Bregman Divergence

Example: For $a \in [0,\infty)^d$, the unnormalized negative entropy, $\Phi(a) = \sum_{i=1}^d a_i\,(\ln a_i - 1)$, has

$$D_\Phi(a,b) = \sum_i (a_i(\ln a_i - 1) - b_i(\ln b_i - 1) - \ln b_i(a_i - b_i))$$

$$= \sum_i \left(a_i \ln \frac{a_i}{b_i} + b_i - a_i \right),$$

the unnormalized KL divergence.

Thus, for $a \in \Delta^d$, $\Phi(a) = \sum_i a_i \ln a_i$ has

$$D_\Phi(a,b) = \sum_i a_i \ln \frac{a_i}{b_i}.$$

Bregman Divergence

When the domain of Φ is $\mathcal{A} \subset \mathbb{R}^d$, in addition to differentiability and strict convexity, we make two more assumptions:

- The interior of $\mathcal{A}$ is convex,
- For a sequence approaching the boundary of $\mathcal{A}$, $\|\nabla\Phi(a_n)\| \to \infty$.

We say that such a Φ is a *Legendre function*.

Bregman Divergence Properties

1. $D_\Phi \geq 0$, $D_\Phi(a, a) = 0$.
2. $D_{A+B} = D_A + D_B$.
3. For ℓ linear, $D_{\Phi+\ell} = D_\Phi$.
4. *Bregman projection*, $\Pi_{\mathcal{A}}^{\Phi}(b) = \arg\min_{a \in \mathcal{A}} D_\Phi(a, b)$ is uniquely defined for closed, convex $\mathcal{A}$.
5. *Generalized Pythagorus*: for closed, convex $\mathcal{A}$, $a^* = \Pi_{\mathcal{A}}^{\Phi}(b)$, $a \in \mathcal{A}$, $D_\Phi(a, b) \geq D_\Phi(a, a^*) + D_\Phi(a^*, b)$.
6. $\nabla_a D_\Phi(a, b) = \nabla\Phi(a) - \nabla\Phi(b)$.
7. For Φ^* the Legendre dual of Φ,

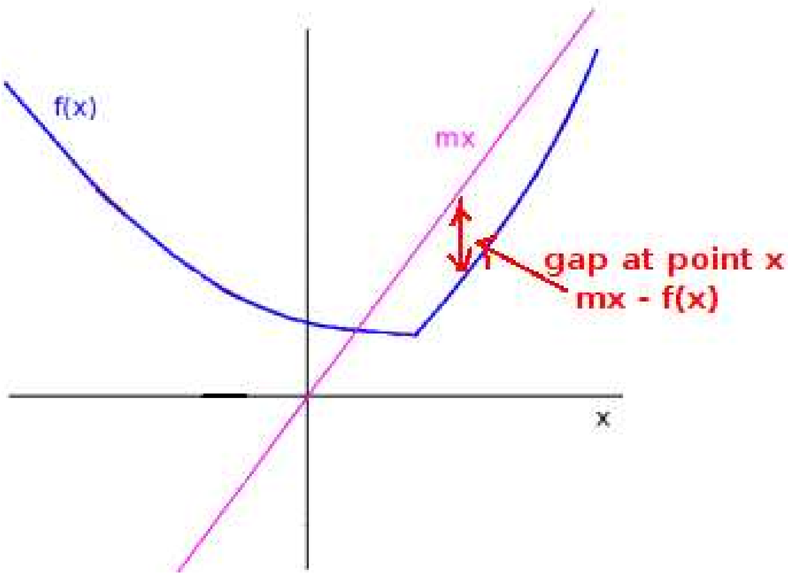
$$\nabla\Phi^* = (\nabla\Phi)^{-1},$$

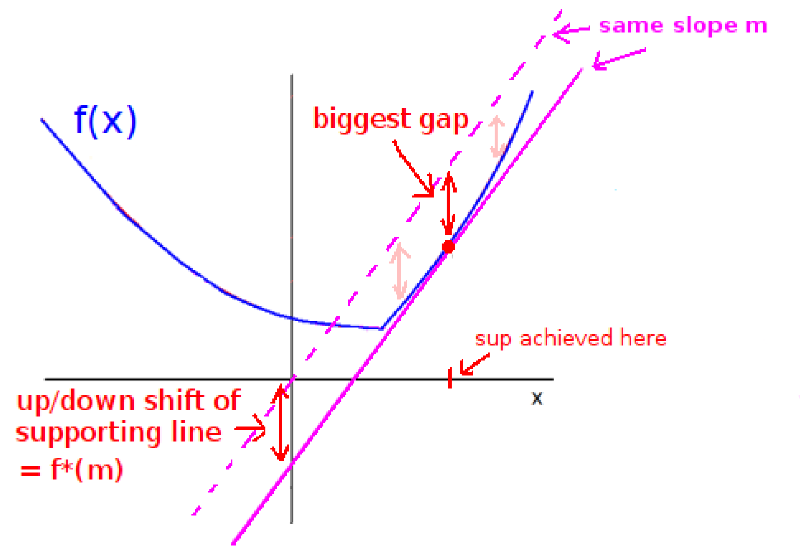
$$D_\Phi(a, b) = D_{\Phi^*}(\nabla\Phi(b), \nabla\Phi(a)).$$

Legendre Dual

Here, for a Legendre function $\Phi : \mathcal{A} \to \mathbb{R}$, we define the Legendre dual as

$$\Phi^*(u) = \sup_{v \in \mathcal{A}} \left(u \cdot v - \Phi(v) \right).$$





(http://maze5.net/)

Legendre Dual

Properties:

- Φ^* is Legendre.
- $\text{dom}(\Phi^*) = \nabla\Phi(\text{int dom } \Phi)$.
- $\nabla\Phi^* = (\nabla\Phi)^{-1}$.
- $D_\Phi(a, b) = D_{\Phi^*}(\nabla\Phi(b), \nabla\Phi(a))$.
- $\Phi^{**} = \Phi$.

Properties of Regularization Methods

In the unconstrained case ($\mathcal{A} = \mathbb{R}^d$), regularized minimization is equivalent to minimizing the latest loss and the distance (Bregman divergence) to the previous decision.

Theorem: Define $\tilde{a}_1$ via $\nabla R(\tilde{a}_1) = 0$, and set

$$\tilde{a}_{t+1} = \arg\min_{a \in \mathbb{R}^d} \left(\eta \ell_t(a) + D_{\Phi_{t-1}}(a, \tilde{a}_t) \right).$$

Then

$$\tilde{a}_{t+1} = \arg\min_{a \in \mathbb{R}^d} \left(\eta \sum_{s=1}^{t} \ell_s(a) + R(a) \right).$$

Properties of Regularization Methods

Proof. By the definition of Φ_t,

$$\eta\ell_t(a) + D_{\Phi_{t-1}}(a, \tilde{a}_t) = \Phi_t(a) - \Phi_{t-1}(a) + D_{\Phi_{t-1}}(a, \tilde{a}_t).$$

The derivative wrt a is

$$\begin{aligned} &\nabla\Phi_t(a) - \nabla\Phi_{t-1}(a) + \nabla_a D_{\Phi_{t-1}}(a, \tilde{a}_t) \\ &= \nabla\Phi_t(a) - \nabla\Phi_{t-1}(a) + \nabla\Phi_{t-1}(a) - \nabla\Phi_{t-1}(\tilde{a}_t) \end{aligned}$$

Setting to zero shows that

$$\nabla\Phi_t(\tilde{a}_{t+1}) = \nabla\Phi_{t-1}(\tilde{a}_t) = \cdots = \nabla\Phi_0(\tilde{a}_1) = \nabla R(\tilde{a}_1) = 0,$$

So $\tilde{a}_{t+1}$ minimizes Φ_t. $\square$

Properties of Regularization Methods

Constrained minimization is equivalent to unconstrained minimization, followed by Bregman projection:

Theorem: For

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} \Phi_t(a),$$

$$\tilde{a}_{t+1} = \arg\min_{a \in \mathbb{R}^d} \Phi_t(a),$$

we have

$$a_{t+1} = \Pi_{\mathcal{A}}^{\Phi_t}(\tilde{a}_{t+1}).$$

Properties of Regularization Methods

Proof. Let a'_{t+1} denote $\Pi_{\mathcal{A}}^{\Phi_t}(\tilde{a}_{t+1})$. First, by definition of a_{t+1},

$$\Phi_t(a_{t+1}) \leq \Phi_t(a'_{t+1}).$$

Conversely,

$$D_{\Phi_t}(a'_{t+1}, \tilde{a}_{t+1}) \leq D_{\Phi_t}(a_{t+1}, \tilde{a}_{t+1}).$$

But $\nabla\Phi_t(\tilde{a}_{t+1}) = 0$, so

$$D_{\Phi_t}(a, \tilde{a}_{t+1}) = \Phi_t(a) - \Phi_t(\tilde{a}_{t+1}).$$

Thus, $\Phi_t(a'_{t+1}) \leq \Phi_t(a_{t+1})$. $\square$

Properties of Regularization Methods

Example: For linear ℓ_t, regularized minimization is equivalent to minimizing the last loss plus the Bregman divergence wrt R to the previous decision:

$$\begin{aligned}
&\arg\min_{a \in \mathcal{A}} \left(\eta \sum_{s=1}^{t} \ell_s(a) + R(a) \right) \\
&= \Pi_{\mathcal{A}}^{R} \left(\arg\min_{a \in \mathbb{R}^d} \left(\eta \ell_t(a) + D_R(a, \tilde{a}_t) \right) \right),
\end{aligned}$$

because adding a linear function to Φ does not change D_Φ.

Properties of Regularization Methods: Linear Loss

We can replace ℓ_t by $\nabla\ell_t(a_t)$, and this leads to an upper bound on regret.

Theorem: Any strategy for online linear optimization, with regret satisfying

$$\sum_{t=1}^{n} g_t \cdot a_t - \min_{a\in\mathcal{A}} \sum_{t=1}^{n} g_t \cdot a \leq C_n(g_1, \ldots, g_n)$$

can be used to construct a strategy for online convex optimization, with regret

$$\sum_{t=1}^{n} \ell_t(a_t) - \min_{a\in\mathcal{A}} \sum_{t=1}^{n} \ell_t(a) \leq C_n(\nabla\ell_1(a_1), \ldots, \nabla\ell_n(a_n)).$$

Proof. Convexity implies $\ell_t(a_t) - \ell_t(a) \leq \nabla\ell_t(a_t) \cdot (a_t - a)$. □

Properties of Regularization Methods: Linear Loss

Key Point:

We can replace ℓ_t by $\nabla \ell_t(a_t)$, and this leads to an upper bound on regret.

Thus, we can work with linear ℓ_t.

Regularization Methods: Mirror Descent

Regularized minimization for linear losses can be viewed as mirror descent—taking a gradient step in a dual space:

Theorem: The decisions

$$\tilde{a}_{t+1} = \arg\min_{a \in \mathbb{R}^d} \left(\eta \sum_{s=1}^{t} g_s \cdot a + R(a) \right)$$

can be written

$$\tilde{a}_{t+1} = (\nabla R)^{-1} (\nabla R(\tilde{a}_t) - \eta g_t) \, .$$

This corresponds to first mapping from $\tilde{a}_t$ through ∇R, then taking a step in the direction $-g_t$, then mapping back through $(\nabla R)^{-1} = \nabla R^*$ to $\tilde{a}_{t+1}$.

Regularization Methods: Mirror Descent

Proof. For the unconstrained minimization, we have

$$\nabla R(\tilde{a}_{t+1}) = -\eta \sum_{s=1}^{t} g_s,$$

$$\nabla R(\tilde{a}_t) = -\eta \sum_{s=1}^{t-1} g_s,$$

so $\nabla R(\tilde{a}_{t+1}) = \nabla R(\tilde{a}_t) - \eta g_t$, which can be written

$$\tilde{a}_{t+1} = \nabla R^{-1} \left(\nabla R(\tilde{a}_t) - \eta g_t \right).$$

□