# CS281B/Stat241B. Statistical Learning Theory. Lecture 17.

**Peter Bartlett**

# 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.

Define $\Phi_0 = R$, $\Phi_t = \Phi_{t-1} + \eta \ell_t$, so that $a_{t+1} = \arg\min_{a \in \mathcal{A}} \Phi_t(a)$.

If we replace $\ell_t$ by $\nabla \ell_t(a_t)$, this leads to an upper bound on regret. Thus, we can assume linear $\ell_t$.

# Recall: Bregman Divergence

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

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

For linear $\ell_t$, $D_{\Phi_t} = D_R$.

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

**Theorem:**

$$\begin{aligned}
\tilde{a}_{t+1} &= \arg\min_{a \in \mathbb{R}^d} \Phi_t(a) \\
&= \arg\min_{a \in \mathbb{R}^d} \left( \eta \ell_t(a) + D_{\Phi_{t-1}}(a, \tilde{a}_t) \right). \\
a_{t+1} &= \arg\min_{a \in \mathcal{A}} \Phi_t(a) \\
&= \Pi_{\mathcal{A}}^{\Phi_t}(\tilde{a}_{t+1}).
\end{aligned}$$

# Recall: Linear Loss

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

Thus, we can work with linear $\ell_t$.

(And then $D_{\Phi_t} = D_R$.)

# 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

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

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

# Online Convex Optimization

1. Problem formulation
2. Empirical minimization fails.
3. Gradient algorithm.
4. Regularized minimization and Bregman divergences
5. Regret bounds
    - Unconstrained minimization
    - Seeing the future
    - Strong convexity
    - Examples (gradient, exponentiated gradient)
    - Extensions

# Regularization Methods: Regret

**Theorem:** For $\mathcal{A} = \mathbb{R}^d$, regularized minimization suffers regret against any $a \in \mathcal{A}$ of

$$\sum_{t=1}^{n} \ell_t(a_t) - \sum_{t=1}^{n} \ell_t(a) = \frac{D_R(a, a_1) - D_{\Phi_n}(a, a_{n+1})}{\eta} + \frac{1}{\eta} \sum_{t=1}^{n} D_{\Phi_t}(a_t, a_{t+1}),$$

and thus

$$\hat{L}_n \leq \inf_{a \in \mathbb{R}^d} \left( \sum_{t=1}^{n} \ell_t(a) + \frac{D_R(a, a_1)}{\eta} \right) + \frac{1}{\eta} \sum_{t=1}^{n} D_{\Phi_t}(a_t, a_{t+1}).$$

So the sizes of the steps $D_{\Phi_t}(a_t, a_{t+1})$ determine the regret bound.

# Regret: Proof

$$D_{\Phi_t}(a, a_{t+1}) = \Phi_t(a) - \left( \Phi_t(a_{t+1}) + \underbrace{\nabla \Phi_t(a_{t+1})}_{=0} \cdot (a - a_{t+1}) \right)$$

$$= \Phi_t(a) - \Phi_t(a_{t+1}).$$

Also, $\quad \eta \ell_t(a) = \Phi_t(a) - \Phi_{t-1}(a).$

$$\begin{aligned}
&\eta \left( \ell_t(a_t) - \ell_t(a) \right) \\
&= \Phi_t(a_t) - \Phi_{t-1}(a_t) - (\Phi_t(a) - \Phi_{t-1}(a)) \\
&= \underbrace{\Phi_t(a_t) - \Phi_t(a_{t+1})} + \underbrace{\Phi_t(a_{t+1}) - \Phi_t(a)} + \underbrace{\Phi_{t-1}(a) - \Phi_{t-1}(a_t)} \\
&= D_{\Phi_t}(a_t, a_{t+1}) - D_{\Phi_t}(a, a_{t+1}) + D_{\Phi_{t-1}}(a, a_t).
\end{aligned}$$

# Regret: Proof

$$
\begin{aligned}
&\eta \sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a)) \\
&= \sum_{t=1}^{n} \big(D_{\Phi_t}(a_t, a_{t+1}) + D_{\Phi_{t-1}}(a, a_t) - D_{\Phi_t}(a, a_{t+1})\big) \\
&= \sum_{t=1}^{n} D_{\Phi_t}(a_t, a_{t+1}) + D_{\Phi_0}(a, a_1) - D_{\Phi_n}(a, a_{n+1}) \\
&= \sum_{t=1}^{n} D_{\Phi_t}(a_t, a_{t+1}) + D_R(a, a_1) - D_{\Phi_n}(a, a_{n+1}) \\
&\leq \sum_{t=1}^{n} D_{\Phi_t}(a_t, a_{t+1}) + D_R(a, a_1).
\end{aligned}
$$

# Regularization Methods: Regret

**Theorem:** For $\mathcal{A} = \mathbb{R}^d$, regularized minimization suffers regret

$$\hat{L}_n \leq \inf_{a \in \mathbb{R}^d} \left( \sum_{t=1}^{n} \ell_t(a) + \frac{D_R(a, a_1)}{\eta} \right) + \frac{1}{\eta} \sum_{t=1}^{n} D_{\Phi_t}(a_t, a_{t+1}).$$

Notice that we can write

$$\begin{aligned} D_{\Phi_t}(a_t, a_{t+1}) &= D_{\Phi_t^*}(\nabla \Phi_t(a_{t+1}), \nabla \Phi_t(a_t)) \\ &= D_{\Phi_t^*}(0, \nabla \Phi_{t-1}(a_t) + \eta \nabla \ell_t(a_t)) \\ &= D_{\Phi_t^*}(0, \eta \nabla \ell_t(a_t)). \end{aligned}$$

So it is the size of the gradient steps, $D_{\Phi_t^*}(0, \eta \nabla \ell_t(a_t))$, that determines the regret.

# Regularization Methods: Regret Bounds

**Example:** Suppose $R = \frac{1}{2}\|\cdot\|^2$. Then we have

$$\hat{L}_n \leq L_n^* + \frac{\|a^* - a_1\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{n}\|g_t\|^2.$$

And if $\|g_t\| \leq G$ and $\|a^* - a_1\| \leq D$, choosing $\eta$ appropriately gives $\hat{L}_n - L_n^* \leq DG\sqrt{n}$.

# Regularization Methods: Regret Bounds

Seeing the future gives small regret:

**Theorem:** For regularized minimization, that is,

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

for all $a \in \mathcal{A}$,

$$\sum_{t=1}^{n} \ell_t(a_{t+1}) - \sum_{t=1}^{n} \ell_t(a) \leq \frac{1}{\eta}(R(a) - R(a_1)).$$

# Regularization Methods: Regret Bounds

*Proof.* Since $a_{t+1}$ minimizes $\Phi_t$,

$$\begin{aligned}
\eta \sum_{s=1}^{t} \ell_s(a) + R(a) &\geq \eta \sum_{s=1}^{t} \ell_s(a_{t+1}) + R(a_{t+1}) \\
&= \eta \ell_t(a_{t+1}) + \eta \sum_{s=1}^{t-1} \ell_s(a_{t+1}) + R(a_{t+1}) \\
&\geq \eta \ell_t(a_{t+1}) + \eta \sum_{s=1}^{t-1} \ell_s(a_t) + R(a_t) \\
&\vdots \\
&\geq \eta \sum_{s=1}^{t} \ell_s(a_{s+1}) + R(a_1).
\end{aligned}$$

## Regularization Methods: Regret Bounds

Thus, if $a_t$ and $a_{t+1}$ are close, then regret is small:

**Corollary:** For all $a \in \mathcal{A}$,

$$\sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a)) \leq \sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a_{t+1})) + \frac{1}{\eta} (R(a) - R(a_1)) .$$

So how can we control the increments $\ell_t(a_t) - \ell_t(a_{t+1})$?

# Regularization Methods: Regret Bounds

**Definition:** We say $R$ is strongly convex wrt a norm $\|\cdot\|$ if, for all $a, b$,

$$R(a) \geq R(b) + \nabla R(b) \cdot (a - b) + \frac{1}{2}\|a - b\|^2.$$

For linear losses and strongly convex regularizers, the dual norm of the gradient is small:

**Theorem:** If $R$ is strongly convex wrt a norm $\|\cdot\|$, and $\ell_t(a) = g_t \cdot a$, then

$$\|a_t - a_{t+1}\| \leq \eta \|g_t\|_*,$$

where $\|\cdot\|_*$ is the dual norm to $\|\cdot\|$:

$$\|v\|_* = \sup\{|v \cdot a| : a \in \mathcal{A}, \|a\| \leq 1\}.$$

# Regularization Methods: Regret Bounds

*Proof.*

$$R(a_t) \geq R(a_{t+1}) + \nabla R(a_{t+1}) \cdot (a_t - a_{t+1}) + \frac{1}{2}\|a_t - a_{t+1}\|^2,$$

$$R(a_{t+1}) \geq R(a_t) + \nabla R(a_t) \cdot (a_{t+1} - a_t) + \frac{1}{2}\|a_t - a_{t+1}\|^2.$$

Combining,

$$\|a_t - a_{t+1}\|^2 \leq (\nabla R(a_t) - \nabla R(a_{t+1})) \cdot (a_t - a_{t+1})$$

Hence,

$$\|a_t - a_{t+1}\| \leq \|\nabla R(a_t) - \nabla R(a_{t+1})\|_* = \|\eta g_t\|_*.$$

□

# Regularization Methods: Regret Bounds

This leads to the regret bound:

**Corollary:** For linear losses, if $R$ is strongly convex wrt $\|\cdot\|$, then for all $a \in \mathcal{A}$,

$$\sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a)) \leq \eta \sum_{t=1}^{n} \|g_t\|_*^2 + \frac{1}{\eta} (R(a) - R(a_1)) .$$

Thus, for $\|g_t\|_* \leq G$ and $R(a) - R(a_1) \leq D^2$, choosing $\eta$ appropriately gives regret no more than $2GD\sqrt{n}$.

# Regularization Methods: Regret Bounds

**Example:** Consider $R(a) = \frac{1}{2}\|a\|^2$, $a_1 = 0$, and $\mathcal{A}$ contained in a Euclidean ball of diameter $D$.
Then $R$ is strongly convex wrt $\|\cdot\|$ and $\|\cdot\|_* = \|\cdot\|$. And the mapping between primal and dual spaces is the identity.
So if $\sup_{a\in\mathcal{A}} \|\nabla \ell_t(a)\| \le G$, then regret is no more than $2GD\sqrt{n}$.

# Regularization Methods: Regret Bounds

**Example:** Consider $\mathcal{A} = \Delta^m$, $R(a) = \sum_i a_i \ln a_i$. Then the mapping between primal and dual spaces is $\nabla R(a) = \ln(a) + 1$ (component-wise). And the divergence is the KL divergence,

$$D_R(a, b) = \sum_i a_i \ln(a_i / b_i).$$

And $R$ is strongly convex wrt $\|\cdot\|_1$.
Suppose that $\|g_t\|_\infty \leq 1$. Also, $R(a) - R(a_1) \leq \ln m$, so the regret is no more than $2\sqrt{n \ln m}$.

# Regularization Methods: Regret Bounds

**Example:** $\mathcal{A} = \Delta^m$, $R(a) = \sum_i a_i \ln a_i$.

What are the updates?

$$
\begin{aligned}
a_{t+1} &= \Pi_{\mathcal{A}}^R(\tilde{a}_{t+1}) \\
&= \Pi_{\mathcal{A}}^R(\nabla R^*(\nabla R(\tilde{a}_t) - \eta g_t)) \\
&= \Pi_{\mathcal{A}}^R(\nabla R^*(\ln(\tilde{a}_t \exp(-\eta g_t)) + 1) \\
&= \Pi_{\mathcal{A}}^R(\tilde{a}_t \exp(-\eta g_t)),
\end{aligned}
$$

where the $\ln$ and $\exp$ functions are applied component-wise.

This is exponentiated gradient: mirror descent with $\nabla R = 1 + \ln$.

It is easy to check that the projection corresponds to normalization, $\Pi_{\mathcal{A}}^R(\tilde{a}) = \tilde{a}/\|a\|_1$.

# Regularization Methods: Regret Bounds

Notice that when the losses are linear, exponentiated gradient is exactly the exponential weights strategy we discussed for a finite comparison class ("experts").

Compare $R(a) = \sum_i a_i \ln a_i$ with $R(a) = \frac{1}{2}\|a\|^2$,
for $\|g_t\|_\infty \leq 1$, $\mathcal{A} = \Delta^m$:

$O(\sqrt{n \ln m})$ versus $O(\sqrt{mn})$.