Learning Methods for Online Prediction Problems

Peter Bartlett
Statistics and EECS
UC Berkeley

Course Synopsis

- A finite comparison class: $\mathcal{A} = \{1, \ldots, m\}$.
  1. "Prediction with expert advice."
  2. With perfect predictions: $\log m$ regret.
  3. Exponential weights strategy: $\sqrt{n \log m}$ regret.
  4. Refinements and extensions.
  5. Statistical prediction with a finite class.
- Converting online to batch.
- Online convex optimization.
- Log loss.

Online to Batch Conversion

- Suppose we have an online strategy that, given observations $\ell_1, \ldots, \ell_{t-1}$, produces $a_t = A(\ell_1, \ldots, \ell_{t-1})$.
- Can we convert this to a method that is suitable for a probabilistic setting? That is, if the ℓ_t are chosen i.i.d., can we use A's choices a_t to come up with a $\hat{a} \in \mathcal{A}$ so that

$$\mathbf{E}\ell_1(\hat{a}) - \min_{a \in \mathcal{A}} \mathbf{E}\ell_1\,(a)$$

  is small?
- Consider the following simple randomized method:
  1. Pick T uniformly from $\{0, \ldots, n\}$.
  2. Let $\hat{a} = A(\ell_{T+1}, \ldots, \ell_n)$.

Online to Batch Conversion

Theorem

If A has a regret bound of C_{n+1} for sequences of length $n + 1$, then for any stationary process generating the $\ell_1, \ldots, \ell_{n+1}$, this method satisfies

$$\mathbf{E}\ell_{n+1}(\hat{a}) - \min_{a \in \mathcal{A}} \mathbf{E}\ell_n(a) \leq \frac{C_{n+1}}{n+1}.$$

(Notice that the expectation averages also over the randomness of the method.)

Online to Batch Conversion

Proof.

$$\begin{aligned}
\mathbf{E}\ell_{n+1}(\hat{a}) &= \mathbf{E}\ell_{n+1}(A(\ell_{T+1},\ldots,\ell_n)) \\
&= \mathbf{E}\frac{1}{n+1}\sum_{t=0}^{n}\ell_{n+1}(A(\ell_{t+1},\ldots,\ell_n)) \\
&= \mathbf{E}\frac{1}{n+1}\sum_{t=0}^{n}\ell_{n-t+1}(A(\ell_1,\ldots,\ell_{n-t})) \\
&= \mathbf{E}\frac{1}{n+1}\sum_{t=1}^{n+1}\ell_t(A(\ell_1,\ldots,\ell_{t-1})) \\
&\le \mathbf{E}\frac{1}{n+1}\left(\min_a\sum_{t=1}^{n+1}\ell_t(a)+C_{n+1}\right) \\
&\le \min_a \mathbf{E}\ell_t(a)+\frac{C_{n+1}}{n+1}.
\end{aligned}$$

Online to Batch Conversion

- The theorem is for the expectation over the randomness of the method.
- For a high probability result, we could
  1. Choose $\hat{a} = \frac{1}{n}\sum_{t=1}^{n} a_t$, provided $\mathcal{A}$ is convex and the ℓ_t are all convex.
  2. Choose

$$\hat{a} = \arg\min_{a_t} \left(\frac{1}{n-t} \sum_{s=t+1}^{n} \ell_s(a_t) + c\sqrt{\frac{\log(n/\delta)}{n-t}} \right).$$

  In both cases, the analysis involves concentration of martingale sequences.
  The second (more general) approach does not recover the C_n/n result: the penalty has the wrong form when $C_n = o(\sqrt{n})$.

Online to Batch Conversion

Key Point:

- An online strategy with regret bound C_n can be converted to a batch method.
  The regret per trial in the probabilistic setting is bounded by the regret per trial in the adversarial setting.

Course Synopsis

- A finite comparison class: $\mathcal{A} = \{1, \ldots, m\}$.
- Converting online to batch.
- Online convex optimization.
  1. Problem formulation
  2. Empirical minimization fails.
  3. Gradient algorithm.
  4. Regularized minimization
  5. Regret bounds
- Log loss.

Online Convex Optimization

1. Problem formulation
2. Empirical minimization fails.
3. Gradient algorithm.
4. Regularized minimization
   - Bregman divergence
   - Regularized minimization equivalent to minimizing latest loss and divergence from previous decision
   - Constrained minimization equivalent to unconstrained plus Bregman projection
   - Linearization
   - Mirror descent
5. Regret bounds
   - Unconstrained minimization
   - Seeing the future
   - Strong convexity
   - Examples (gradient, exponentiated gradient)
   - Extensions

Online Convex Optimization

- $\mathcal{A} =$ convex subset of $\mathbb{R}^d$.
- $\mathcal{L} =$ set of convex real functions on $\mathcal{A}$.

For example,

$$\ell_t(a) = (x_t \cdot a - y_t)^2.$$

$$\ell_t(a) = |x_t \cdot a - y_t|.$$

Online Convex Optimization: Example

Choosing a_t to minimize past losses, $a_t = \arg\min_{a \in \mathcal{A}} \sum_{s=1}^{t-1} \ell_s(a)$, can fail. ('fictitious play,' 'follow the leader')

- Suppose $\mathcal{A} = [-1, 1]$, $\mathcal{L} = \{a \mapsto v \cdot a : |v| \le 1\}$.
- Consider the following sequence of losses:

$$
\begin{aligned}
a_1 &= 0, & \ell_1(a) &= \frac{1}{2}a, \\
a_2 &= -1, & \ell_2(a) &= -a, \\
a_3 &= 1, & \ell_3(a) &= a, \\
a_4 &= -1, & \ell_4(a) &= -a, \\
a_5 &= 1, & \ell_5(a) &= a, \\
&\vdots & &\vdots
\end{aligned}
$$

- $a^* = 0$ shows $L_n^* \le 0$, but $\hat{L}_n = n - 1$.

Online Convex Optimization: Example

- Choosing a_t to minimize past losses can fail.
- The strategy must avoid overfitting, just as in probabilistic settings.
- Similar approaches (regularization; Bayesian inference) are applicable in the online setting.
- First approach: gradient steps.
  Stay close to previous decisions, but move in a direction of improvement.

Online Convex Optimization: Gradient Method

$$a_1 \in \mathcal{A},$$
$$a_{t+1} = \Pi_{\mathcal{A}}\left(a_t - \eta \nabla \ell_t(a_t)\right),$$

where $\Pi_{\mathcal{A}}$ is the Euclidean projection on $\mathcal{A}$,

$$\Pi_{\mathcal{A}}(x) = \arg\min_{a \in \mathcal{A}} \|x - a\|.$$

Theorem

For $G = \max_t \|\nabla \ell_t(a_t)\|$ and $D = diam(\mathcal{A})$, the gradient strategy with $\eta = D/(G\sqrt{n})$ has regret satisfying

$$\hat{L}_n - L_n^* \leq GD\sqrt{n}.$$

Online Convex Optimization: Gradient Method

Theorem

For $G = \max_t \|\nabla \ell_t(a_t)\|$ *and* $D = diam(\mathcal{A})$, *the gradient strategy with* $\eta = D/(G\sqrt{n})$ *has regret satisfying*

$$\hat{L}_n - L_n^* \leq GD\sqrt{n}.$$

Example

$\mathcal{A} = \{a \in \mathbb{R}^d : \|a\| \leq 1\}$, $\mathcal{L} = \{a \mapsto v \cdot a : \|v\| \leq 1\}$.
$D = 2$, $G \leq 1$.
Regret is no more than $2\sqrt{n}$.

(And $O(\sqrt{n})$ is optimal.)

Online Convex Optimization: Gradient Method

Theorem

For $G = \max_t \|\nabla \ell_t(a_t)\|$ and $D = diam(\mathcal{A})$, the gradient strategy with $\eta = D/(G\sqrt{n})$ has regret satisfying

$$\hat{L}_n - L_n^* \leq GD\sqrt{n}.$$

Example

$\mathcal{A} = \Delta^m$, $\mathcal{L} = \{a \mapsto v \cdot a : \|v\|_\infty \leq 1\}$.
$D = 2$, $G \leq \sqrt{m}$.
Regret is no more than $2\sqrt{mn}$.

Since competing with the whole simplex is equivalent to competing with the vertices (experts) for linear losses, this is worse than exponential weights ($\sqrt{m}$ versus $\log m$).

Online Convex Optimization: Gradient Method

Proof.

Define
$$\tilde{a}_{t+1} = a_t - \eta \nabla \ell_t(a_t),$$
$$a_{t+1} = \Pi_{\mathcal{A}}(\tilde{a}_{t+1}).$$

Fix $a \in \mathcal{A}$ and consider the measure of progress $\|a_t - a\|$.

$$\begin{aligned}\|a_{t+1} - a\|^2 &\le \|\tilde{a}_{t+1} - a\|^2 \\ &= \|a_t - a\|^2 + \eta^2 \|\nabla \ell_t(a_t)\|^2 - 2\eta \nabla_t(a_t) \cdot (a_t - a).\end{aligned}$$

By convexity,

$$\begin{aligned}\sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a)) &\le \sum_{t=1}^{n} \nabla \ell_t(a_t) \cdot (a_t - a) \\ &\le \frac{\|a_1 - a\|^2 - \|a_{n+1} - a\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \|\nabla \ell_t(a_t)\|^2\end{aligned}$$

Online Convex Optimization

1. Problem formulation
2. Empirical minimization fails.
3. Gradient algorithm.
4. Regularized minimization
   - Bregman divergence
   - Regularized minimization equivalent to minimizing latest loss and divergence from previous decision
   - Constrained minimization equivalent to unconstrained plus Bregman projection
   - Linearization
   - Mirror descent
5. Regret bounds
   - Unconstrained minimization
   - Seeing the future
   - Strong convexity
   - Examples (gradient, exponentiated gradient)
   - Extensions

Online Convex Optimization: A Regularization Viewpoint

- Suppose ℓ_t is linear: $\ell_t(a) = g_t \cdot a$.
- Suppose $\mathcal{A} = \mathbb{R}^d$.
- Then minimizing the regularized criterion

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

corresponds to the gradient step

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

Online Convex Optimization: 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.

Online Convex Optimization: Regularization

Regularized minimization

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

- 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.
- This is a perspective that motivated many algorithms in the literature. We'll investigate why regularized minimization can be viewed this way.

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 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

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

Bregman Divergence

Definition

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.

Bregman Divergence

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

Example

For $a \in \mathbb{R}^d$, the squared euclidean norm, $\Phi(a) = \frac{1}{2}\|a\|^2$, has

$$D_\Phi(a,b) = \frac{1}{2}\|a\|^2 - \left(\frac{1}{2}\|b\|^2 + b\cdot(a-b)\right)$$
$$= \frac{1}{2}\|a-b\|^2,$$

the squared euclidean norm.

Bregman Divergence

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

Example

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

$$\begin{aligned} D_\Phi(a, b) &= \sum_i (a_i(\ln a_i - 1) - b_i(\ln b_i - 1) - \ln b_i(a_i - b_i)) \\ &= \sum_i \left(a_i \ln \frac{a_i}{b_i} + b_i - a_i \right) , \end{aligned}$$

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 range 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. *Bregman projection*, $\Pi^\Phi_{\mathcal{A}}(b) = \arg\min_{a \in \mathcal{A}} D_\Phi(a,b)$ is uniquely defined for closed, convex $\mathcal{A}$.
4. *Generalized Pythagorus*: for closed, convex $\mathcal{A}$, $b^* = \Pi^\Phi_{\mathcal{A}}(b)$, and $a \in \mathcal{A}$,
$$D_\Phi(a,b) \geq D_\Phi(a,a^*) + D_\Phi(a^*,b).$$
5. $\nabla_a D_\Phi(a,b) = \nabla\Phi(a) - \nabla\Phi(b)$.
6. For ℓ linear, $D_{\Phi+\ell} = D_\Phi$.
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

For a Legendre function $\Phi : \mathcal{A} \to \mathbb{R}$, the Legendre dual is

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

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

Legendre Dual

Example

For $\Phi = \frac{1}{2}\|\cdot\|_p^2$, the Legendre dual is $\Phi^* = \frac{1}{2}\|\cdot\|_q^2$, where $1/p + 1/q = 1$.

Example

For $\Phi(a) = \sum_{i=1}^{d} e^{a_i}$,

$$\nabla\Phi(a) = (e^{a_1}, \ldots, e^{a_d})',$$

so

$$(\nabla\Phi)^{-1}(u) = \nabla\Phi^*(u) = (\ln u_1, \ldots, \ln u_d)',$$

and $\Phi^*(u) = \sum_i u_i(\ln u_i - 1)$.

Online Convex Optimization

1. Problem formulation
2. Empirical minimization fails.
3. Gradient algorithm.
4. Regularized minimization
   - Bregman divergence
   - Regularized minimization equivalent to minimizing latest loss and divergence from previous decision
   - Constrained minimization equivalent to unconstrained plus Bregman projection
   - Linearization
   - Mirror descent
5. Regret bounds
   - Unconstrained minimization
   - Seeing the future
   - Strong convexity
   - Examples (gradient, exponentiated gradient)
   - Extensions

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

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:

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

$$= \Pi_{\mathcal{A}}^{R} \left(\arg\min_{a \in \mathbb{R}^d} \left(\eta \ell_t(a) + D_R(a, \tilde{a}_t) \right) \right),$$

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

Online Convex Optimization

1. Problem formulation
2. Empirical minimization fails.
3. Gradient algorithm.
4. Regularized minimization
    - Bregman divergence
    - Regularized minimization equivalent and Bregman divergence from previous
    - Constrained minimization equivalent to unconstrained plus Bregman projection
    - Linearization
    - Mirror descent
5. Regret bounds
    - Unconstrained minimization
    - Seeing the future
    - Strong convexity
    - Examples (gradient, exponentiated gradient)
    - Extensions

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 $\ell_t(a_t) - \ell_t(a) \leq \nabla \ell_t(a_t) \cdot (a_t - a)$. $\square$

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

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

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

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

□

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

Online Convex Optimization: Regularization

Regularized minimization

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

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

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.

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 η appropriately gives $\hat{L}_n \leq L_n^* \leq DG\sqrt{n}$.

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 Bounds

Seeing the future gives small regret:

Theorem

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 Φ_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

Theorem

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

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

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 $\ell_t(a_t) - \ell_t(a_{t+1})$?

Online Convex Optimization

1. Problem formulation
2. Empirical minimization fails.
3. Gradient algorithm.
4. Regularized minimization
   - Bregman divergence
   - Regularized minimization equivalent and Bregman divergence from previous
   - Constrained minimization equivalent to unconstrained plus Bregman projection
   - Linearization
   - Mirror descent
5. Regret bounds
   - Unconstrained minimization
   - Seeing the future
   - Strong convexity
   - Examples (gradient, exponentiated gradient)
   - Extensions

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 η 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 $\sup_{a \in \mathcal{A}} \|\nabla \ell_t(a)\| \leq G$, 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 $\nabla R(a) = \ln(a)$ (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 $\|\cdot\|_1$ (check!).
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?

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

where the ln and exp functions are applied component-wise.
This is exponentiated gradient: mirror descent with $\nabla R = \ln$.
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.
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 $O(\sqrt{mn})$.

Online Convex Optimization

1. Problem formulation
2. Empirical minimization fails.
3. Gradient algorithm.
4. Regularized minimization
   - Bregman divergence
   - Regularized minimization equivalent and Bregman divergence from previous
   - Constrained minimization equivalent to unconstrained plus Bregman projection
   - Linearization
   - Mirror descent
5. Regret bounds
   - Unconstrained minimization
   - Strong convexity
   - Examples (gradient, exponentiated gradient)
   - Extensions

Regularization Methods: Extensions

- Instead of

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} \left(\eta \ell_t(a) + D_{\Phi_{t-1}}(a, \tilde{a}_t) \right),$$

  we can use

$$a_{t+1} = \arg\min_{a \in \mathcal{A}} \left(\eta \ell_t(a) + D_{\Phi_{t-1}}(a, a_t) \right).$$

  And analogous results apply. For instance, this is the approach used by the first gradient method we considered.
- We can get faster rates with stronger assumptions on the losses...

Regularization Methods: Varying η

Theorem

Define

$$a_{t+1} = \arg\min_{a \in \mathbb{R}^d} \left(\sum_{t=1}^{n} \eta_t \ell_t(a) + R(a) \right).$$

For any $a \in \mathbb{R}^d$,

$$\hat{L}_n - \sum_{t=1}^{n} \ell_t(a) \leq \sum_{t=1}^{n} \frac{1}{\eta_t} \left(D_{\Phi_t}(a_t, a_{t+1}) + D_{\Phi_{t-1}}(a, a_t) - D_{\Phi_t}(a, a_{t+1}) \right).$$

If we linearize the ℓ_t, we have

$$\hat{L}_n - \sum_{t=1}^{n} \ell_t(a) \leq \sum_{t=1}^{n} \frac{1}{\eta_t} \left(D_R(a_t, a_{t+1}) + D_R(a, a_t) - D_R(a, a_{t+1}) \right).$$

But what if ℓ_t are strongly convex?

Regularization Methods: Strongly Convex Losses

Theorem

If ℓ_t is σ-strongly convex wrt R, that is, for all $a, b \in \mathbb{R}^d$,

$$\ell_t(a) \geq \ell_t(b) + \nabla \ell_t(b) \cdot (a - b) + \frac{\sigma}{2} D_R(a, b),$$

then for any $a \in \mathbb{R}^d$, this strategy with $\eta_t = \frac{2}{t\sigma}$ has regret

$$\hat{L}_n - \sum_{t=1}^{n} \ell_t(a) \leq \sum_{t=1}^{n} \frac{1}{\eta_t} D_R(a_t, a_{t+1}).$$

Strongly Convex Losses: Proof idea

$$
\begin{aligned}
&\sum_{t=1}^{n} (\ell_t(a_t) - \ell_t(a)) \\
&\le \sum_{t=1}^{n} \left(\nabla \ell_t(a_t) \cdot (a_t - a) - \frac{\sigma}{2} D_R(a, a_t) \right) \\
&\le \sum_{t=1}^{n} \frac{1}{\eta_t} \left(D_R(a_t, a_{t+1}) + D_R(a, a_t) - D_R(a, a_{t+1}) - \frac{\eta_t \sigma}{2} D_R(a, a_t) \right) \\
&\le \sum_{t=1}^{n} \frac{1}{\eta_t} D_R(a_t, a_{t+1}) + \sum_{t=2}^{n} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \frac{\sigma}{2} \right) D_R(a, a_t) \\
&\qquad + \left(\frac{1}{\eta_1} - \frac{\sigma}{2} \right) D_R(a, a_1).
\end{aligned}
$$

And choosing η_t appropriately eliminates the second and third terms.

Strongly Convex Losses

Example

For $R(a) = \frac{1}{2}\|a\|^2$, we have

$$\hat{L}_n - L_n^* \leq \frac{1}{2}\sum_{t=1}^{n}\frac{1}{\eta_t}\|\eta_t \nabla \ell_t\|^2 \leq \sum_{t=1}^{n}\frac{G^2}{t\sigma} = O\left(\frac{G^2}{\sigma}\log n\right).$$

Strongly Convex Losses

Key Point: When the loss is strongly convex wrt the regularizer, the regret rate can be faster; in the case of quadratic R (and ℓ_t), it is $O(\log n)$, versus $O(\sqrt{n})$.

Course Synopsis

- A finite comparison class: $\mathcal{A} = \{1, \ldots, m\}$.
- Converting online to batch.
- Online convex optimization.
- Log loss.