CMPT 409/981: Optimization for Machine Learning

Lecture 15

Sharan Vaswani

October 31, 2024

Recap: Online Optimization

Generic Online Optimization (w_0, Algorithm $\mathcal{A}$, Convex set $\mathcal{C} \subseteq \mathbb{R}^d$)

for $k = 1, \ldots, T$ **do**
  Algorithm $\mathcal{A}$ chooses point (decision) $w_k \in \mathcal{C}$
  Environment chooses and reveals the (potentially adversarial) loss function $f_k : \mathcal{C} \to \mathbb{R}$
  Algorithm suffers a cost $f_k(w_k)$
end for

Application: **Prediction from Expert Advice**: Given d experts, $\mathcal{C} = \Delta_d = \{w_i | w_i \geq 0\ ;\ \sum_{i=1}^{d} w_i = 1\}$ and $f_k(w_k) = \langle c_k, w_k \rangle$ where $c_k \in \mathbb{R}^d$ is the loss vector.

Application: **Imitation Learning**: Given access to an expert that knows what action $a \in [A]$ to take in each state $s \in [S]$, learn a policy $\pi : [S] \to [A]$ that imitates the expert, i.e. we want that $\pi(a|s) \approx \pi_{\text{expert}}(a|s)$. Here, $w = \pi$ and $\mathcal{C} = \Delta_A \times \Delta_A \ldots \Delta_A$ (simplex for each state) and f_k is a measure of discrepancy between π_k and π_{expert}.

Online Optimization

- Recall that the sequence of losses $\{f_k\}_{k=1}^{T}$ is potentially adversarial and can also depend on w_k.
- **Objective**: Do well against the *best fixed decision in hindsight*, i.e. if we knew the entire sequence of losses beforehand, we would choose $w^* := \arg\min_{w \in \mathcal{C}} \sum_{k=1}^{T} f_k(w)$.
- **Regret**: For any fixed decision $u \in \mathcal{C}$,

$$R_T(u) := \sum_{k=1}^{T} [f_k(w_k) - f_k(u)]$$

When comparing against the best decision in hindsight,

$$R_T := \sum_{k=1}^{T} [f_k(w_k)] - \min_{w \in \mathcal{C}} \sum_{k=1}^{T} f_k(w).$$

- We want to design algorithms that achieve a *sublinear regret* (that grows as $o(T)$). A sublinear regret implies that the performance of our sequence of decisions is approaching that of w^*.

Online Convex Optimization

- **Online Convex Optimization** (OCO): When the losses f_k are (strongly) convex loss functions.

Example 1: In prediction with expert advice, $f_k(w) = \langle c_k, w \rangle$ is a linear function.

Example 2: In imitation learning, $f_k(\pi) = \mathbb{E}_{s \sim d^{\pi_k}}[\mathrm{KL}(\pi(\cdot|s) \,||\, \pi_{\text{expert}}(\cdot|s)]$ where d^{π_k} is a distribution over the states induced by running policy π_k.

Example 3: In online control such as LQR (linear quadratic regulator) with unknown costs/perturbations, f_k is quadratic.

- In Examples 2-3, the loss at iteration $k+1$ depends on the *learner*'s decision at iteration k.

Online Convex Optimization

• **Online-to-Batch conversion**: If the sequence of loss functions is i.i.d from some fixed distribution, we can convert the regret guarantees into the traditional convergence guarantees for the resulting algorithm.

Formally, if f_k are convex and $R(T) = O(\sqrt{T})$, then taking the expectation w.r.t the distribution generating the losses,

$$\mathbb{E}\left[\frac{R_T}{T}\right] = \mathbb{E}\left[\frac{\sum_{k=1}^{T}[f_k(w_k)] - \sum_{k=1}^{T} f_k(w^*)}{T}\right] \geq \sum_{k=1}^{T}[f(\bar{w}_T) - f(w^*)] = O\left(\frac{1}{\sqrt{T}}\right)$$

where $f(w) := \mathbb{E}[f_k(w)]$ (since the losses are i.i.d) and $\bar{w}_T := \frac{\sum_{k=1}^{T} w_k}{T}$ (since the losses are convex, we used Jensen's inequality).

• If the distribution generating the losses is a uniform discrete distribution on n fixed data-points, then $f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$ and we are back in the finite-sum minimization setting.

• Hence, algorithms that attain $R(T) = O(\sqrt{T})$ can result in an $O\left(\frac{1}{\sqrt{T}}\right)$ convergence (in terms of the function values) for convex losses.

Questions?

Online Gradient Descent

The simplest algorithm that results in sublinear regret for OCO is *Online Gradient Descent*.

Online Gradient Descent (OGD): At iteration k, the algorithm chooses the point w_k. After the loss function f_k is revealed, OGD suffers a cost $f_k(w_k)$ and uses the function to compute

$$w_{k+1} = \Pi_C[w_k - \eta_k \nabla f_k(w_k)]$$

where $\Pi_C[x] = \arg\min_{y \in \mathcal{C}} \frac{1}{2} \|y - x\|^2$.

Claim: If the convex set $\mathcal{C}$ has a diameter D i.e. for all $x, y \in \mathcal{C}$, $\|x - y\| \leq D$, for an arbitrary sequence of losses such that each f_k is convex and differentiable, OGD with a non-increasing sequence of step-sizes i.e. $\eta_k \leq \eta_{k-1}$ and $w_1 \in \mathcal{C}$ has the following regret for all $u \in \mathcal{C}$,

$$R_T(u) \leq \frac{D^2}{2\eta_T} + \sum_{k=1}^{T} \frac{\eta_k}{2} \|\nabla f_k(w_k)\|^2$$

Online Gradient Descent - Convex functions

Proof: Using the update $w_{k+1} = \Pi_{\mathcal{C}}[w_k - \eta_k \nabla f_k(w_k)]$. Since $u \in \mathcal{C}$,

$$\|w_{k+1} - u\|^2 = \|\Pi_{\mathcal{C}}[w_k - \eta_k \nabla f_k(w_k)] - u\|^2 = \|\Pi_{\mathcal{C}}[w_k - \eta_k \nabla f_k(w_k)] - \Pi_{\mathcal{C}}[u]\|^2$$

Since projections are non-expansive i.e. for all x, y, $\|\Pi_{\mathcal{C}}[y] - \Pi_{\mathcal{C}}[x]\| \leq \|y - x\|$,

$$\begin{aligned}
&\leq \|w_k - \eta_k \nabla f_k(w_k) - u\|^2 \\
&= \|w_k - u\|^2 - 2\eta_k \langle \nabla f_k(w_k), w_k - u \rangle + \eta_k^2 \|\nabla f_k(w_k)\|^2 \\
&\leq \|w_k - u\|^2 - 2\eta_k [f_k(w_k) - f_k(u)] + \eta_k^2 \|\nabla f_k(w_k)\|^2
\end{aligned}$$

(Since f_k is convex)

$$\implies 2\eta_k [f_k(w_k) - f_k(u)] \leq [\|w_k - u\|^2 - \|w_{k+1} - u\|^2] + \eta_k^2 \|\nabla f_k(w_k)\|^2$$

$$\implies R_T(u) \leq \sum_{k=1}^{T} \left[\frac{\|w_k - u\|^2 - \|w_{k+1} - u\|^2}{2\eta_k} \right] + \sum_{k=1}^{T} \frac{\eta_k}{2} \|\nabla f_k(w_k)\|^2$$

Online Gradient Descent - Convex functions

Recall that $R_T(u) \leq \sum_{k=1}^{T} \left[\frac{\|w_k - u\|^2 - \|w_{k+1} - u\|^2}{2\eta_k}\right] + \sum_{k=1}^{T} \frac{\eta_k}{2} \|\nabla f_k(w_k)\|^2$.

$$\sum_{k=1}^{T} \left[\frac{\|w_k - u\|^2 - \|w_{k+1} - u\|^2}{2\eta_k}\right]$$

$$= \sum_{k=2}^{T} \left[\|w_k - u\|^2 \quad \underbrace{\left(\frac{1}{2\eta_k} - \frac{1}{2\eta_{k-1}}\right)}_{\text{Non-negative since } \eta_k \leq \eta_{k-1}}\right] + \frac{\|w_1 - u\|^2}{2\eta_1} - \frac{\|w_{T+1} - u\|^2}{2\eta_T}$$

$$\leq D^2 \sum_{k=2}^{T} \left[\frac{1}{2\eta_k} - \frac{1}{2\eta_{k-1}}\right] + \frac{D^2}{2\eta_1} = D^2 \left[\frac{1}{2\eta_T} - \frac{1}{2\eta_1}\right] + \frac{D^2}{2\eta_1} = \frac{D^2}{2\eta_T}$$

(Since $\|x - y\| \leq D$ for all $x, y \in \mathcal{C}$)

Putting everything together,

$$R_T(u) \leq \frac{D^2}{2\eta_T} + \sum_{k=1}^{T} \frac{\eta_k}{2} \|\nabla f_k(w_k)\|^2$$

Online Gradient Descent - Convex, Lipschitz functions

Claim: If the convex set $\mathcal{C}$ has a diameter D i.e. for all $x, y \in \mathcal{C}$, $\|x - y\| \leq D$, for an arbitrary sequence of losses such that each f_k is convex, differentiable and G-Lipschitz, OGD with $\eta_k = \frac{\eta}{\sqrt{k}}$ and $w_1 \in \mathcal{C}$ has the following regret for all $u \in \mathcal{C}$,

$$R_T(u) \leq \frac{D^2 \sqrt{T}}{2\eta} + G^2 \sqrt{T}\, \eta$$

Proof: Since the step-size is decreasing, we can use the general result from the previous slide,

$$R_T(u) \leq \frac{D^2}{2\eta_T} + \sum_{k=1}^{T} \frac{\eta_k}{2} \left\|\nabla f_k(w_k)\right\|^2 \leq \frac{D^2}{2\eta_T} + \frac{G^2}{2} \sum_{k=1}^{T} \eta_k \qquad \text{(Since } f_k \text{ is } G\text{-Lipschitz)}$$

$$\implies R_T(u) \leq \frac{D^2 \sqrt{T}}{2\eta} + \frac{G^2 \eta}{2} \sum_{k=1}^{T} \frac{1}{\sqrt{k}} \leq \frac{D^2 \sqrt{T}}{2\eta} + G^2 \sqrt{T}\, \eta \qquad \left(\text{Since } \textstyle\sum_{k=1}^{T} \frac{1}{\sqrt{k}} \leq 2\sqrt{T}\right)$$

• In order to find the "best" η, set it such that $D^2/2\eta = G^2 \eta$, implying that $\eta = D/\sqrt{2}G$ and $R_T(u) \leq \sqrt{2}\, DG \sqrt{T}$. Hence, OGD with a decreasing step-size attains sublinear $\Theta(\sqrt{T})$ regret for convex, Lipschitz functions.

Questions?

Online Mirror Descent

- The OGD update at iteration k can also be written as:
$w_{k+1} = \arg\min_{w \in \mathcal{C}} \left[\langle \nabla f_k(w_k), w \rangle + \frac{1}{2\eta_k} \, \|w - w_k\|_2^2 \right]$

- Online Mirror Descent (OMD) generalizes gradient descent by choosing a strictly convex, differentiable function $\phi : \mathbb{R}^d \to \mathbb{R}$ (referred to as the *mirror map*) to induce a distance measure.

- ϕ induces the *Bregman divergence* $D_\phi(\cdot, \cdot)$, a distance measure between points x, y,

$$D_\phi(y, x) := \phi(y) - \phi(x) - \langle \nabla \phi(x), y - x \rangle \, .$$

Geometrically, $D_\phi(y, x)$ is the distance between the function $\phi(y)$ and the line $\phi(x) + \langle \nabla \phi(x), y - x \rangle$ which is tangent to the function at x.

- Using D_ϕ as the distance measure results in the mirror descent update:

$$w_{k+1} = \arg\min_{w \in \mathcal{C}} \left[\langle \nabla f_k(w_k), w \rangle + \frac{1}{\eta_k} D_\phi(w, w_k) \right]$$

- Setting $\phi(x) = \frac{1}{2} \|x\|^2$ results in $D_\phi(y, x) = \frac{1}{2} \|y - x\|^2$ and recovers OGD.

Online Mirror Descent – Example

- For prediction with expert advice, $\mathcal{C} = \Delta_d = \{w_i | w_i \geq 0 \; ; \; \sum_{i=1}^{d} w_i = 1\}$ and we want a distance metric between probabilities.
- Typically use the *negative-entropy mirror map* i.e. $\phi(w) = \sum_{i=1}^{d} w_i \ln(w_i)$.
- For $u, v \in \mathcal{C}$, the corresponding Bregman divergence $D_\phi(u, v)$ can be calculated as:

$$D_\phi(u, v) = \phi(u) - \phi(v) - \langle \nabla\phi(v), u - v\rangle = \phi(u) - \phi(v) - \langle \log(v) + 1, u - v\rangle$$

$$(\nabla\phi(u) = \log(u) + 1, \text{ where } \log(\cdot) \text{ is element-wise})$$

$$= \sum_{i=1}^{d} u_i \log(u_i) - \sum_{i=1}^{d} v_i \log(v_i) - \left[\sum_{i=1}^{d} u_i \log(v_i) - \sum_{i=1}^{d} v_i \log(v_i)\right] - \sum_{i=1}^{d} (u_i - v_i)$$

$$= \sum_{i=1}^{d} u_i \log\left(\frac{u_i}{v_i}\right) = \mathrm{KL}(u||v). \qquad \left(\textstyle\sum_{i=1}^{d} u_i = \sum_{i=1}^{d} v_i = 1\right)$$

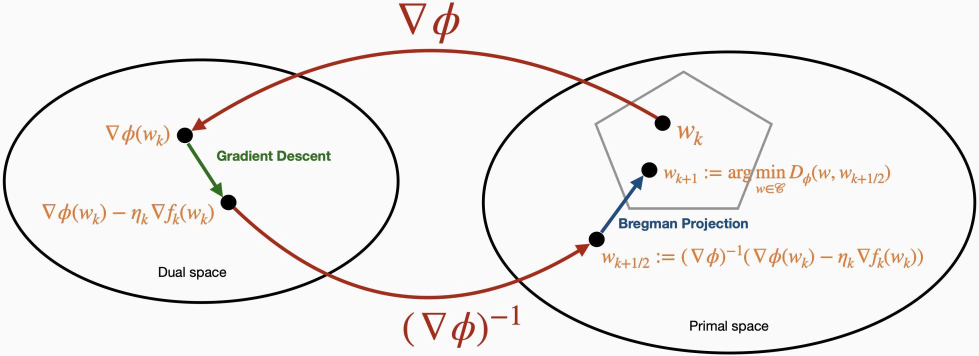
- The KL-divergence is a standard way to measure the distance between probability distributions. For distributions u, v, $\mathrm{KL}(u||v) := \sum_{i=1}^{d} u_i \log\left(\frac{u_i}{v_i}\right)$ is non-negative and equal to zero iff $u = v$.

Online Mirror Descent

The OMD update can be equivalently written as:

GD in dual space: $w_{k+1/2} = (\nabla\phi)^{-1}\left(\nabla\phi(w_k) - \eta_k \nabla f_k(w_k)\right)$

Bregman projection: $w_{k+1} = \arg\min_{w\in\mathcal{C}} D_\phi(w, w_{k+1/2})$



Prove in Assignment 3!

Online Mirror Descent – Example

For prediction with expert advice, $\mathcal{C} = \Delta_d = \{w_i | w_i \geq 0 \; ; \; \sum_{i=1}^{d} w_i = 1\}$, $\phi(w) = \sum_{i=1}^{d} w_i \ln(w_i)$ is the negative-entropy mirror map and $g_k := \nabla f_k(w_k)$, then the OMD update can be written as: (prove in Assignment 3!)

- **GD in dual space**: $w_{k+1/2}[i] = w_k[i] \exp(-\eta_k \, g_k[i])$
- **Bregman projection**: $w_{k+1}[i] = \frac{w_{k+1/2}[i]}{\left\| w_{k+1/2} \right\|_1}$

- **Multiplicative weights update:**

$$w_{k+1}[i] = \frac{w_k[i] \, \exp\left(-\eta_k \, g_k[i]\right)}{\sum_{j=1}^{d} w_k[j] \, \exp\left(-\eta_k \, g_k[j]\right)}$$

If $w_0[i] = \frac{1}{d}$ for all $i \in [d]$, then, for all k,

$$w_{k+1}[i] = \frac{\exp\left(-\sum_{m=1}^{k} \eta_m \, g_m[i]\right)}{\sum_{j=1}^{d} \exp\left(-\sum_{m=1}^{k} \eta_m \, g_m[j]\right)}$$

Online Mirror Descent – Convex, Lipschitz functions

In order to analyze OMD, we will make some assumptions about $\mathcal{C}$, f_k and ϕ.

- **Assumption 1**: $\mathcal{C}$ is a convex set and $\forall k$, f_k is a convex function.
- **Assumption 2**: $\forall k$, f_k is G-Lipschitz in the ℓ_p norm (for $p \geq 1$), implying that $\forall w \in \mathcal{C}$,

$$\left\|\nabla f_k(w)\right\|_p \leq G$$

- **Assumption 3**: ϕ is ν strongly-convex in the ℓ_q norm (for $q \geq 1$ s.t. $\frac{1}{p} + \frac{1}{q} = 1$) i.e.

$$\phi(y) \geq \phi(x) + \langle \nabla \phi(x), y - x \rangle + \frac{\nu}{2} \left\|y - x\right\|_q^2$$

  - *Example*: For prediction from expert advice,
    - $\mathcal{C} = \Delta_d$ is a convex set and $f_k(w_k) = \langle c_k, w_k \rangle$ is a convex function.
    - If the costs are bounded by M, then, $\left\|\nabla f_k(w)\right\|_\infty = \left\|c_k\right\|_\infty \leq M$. Hence, $p = \infty$, $G = M$.
    - If $\phi(w)$ is negative-entropy, then by Pinsker's inequality, $q = 1$ and $\nu = 1$ i.e.

$$\phi(y) - \phi(x) - \langle \nabla \phi(x), y - x \rangle = D_\phi(y, x) = \mathrm{KL}(y||x) \geq \frac{1}{2} \left\|y - x\right\|_1^2 .$$