

Advanced Machine Learning

Online Convex Optimization

MEHRYAR MOHRI

MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

Outline

- Online projected sub-gradient descent.
- Exponentiated Gradient (EG).
- Mirror descent.
- Dual Averaging.

Set-Up

- Convex set C .
- For $t = 1$ to T do
 - predict $\mathbf{w}_t \in C$.
 - receive convex loss function $f_t : C \rightarrow \mathbb{R}$.
 - incur loss $f_t(\mathbf{w}_t)$.
- Regret of algorithm \mathcal{A} :

$$R_T(\mathcal{A}) = \sum_{t=1}^T f_t(\mathbf{w}_t) - \inf_{\mathbf{w} \in C} \sum_{t=1}^T f_t(\mathbf{w}).$$

Online Projected Subgrad. Desc.

■ Algorithm:

- $\mathbf{w}_1 \in C$ arbitrary.
- $\mathbf{w}_{t+1} = \Pi_C[\mathbf{w}_t - \eta \delta f_t(\mathbf{w}_t)]$, where
 - Π_C is the projection over C .
 - $\delta f_t(\mathbf{w}_t) \in \partial f_t(\mathbf{w}_t)$ (sub-gradient of f_t at \mathbf{w}_t).
 - $\eta > 0$ parameter.

Analysis

(Zinkevich, 2009)

■ Assumptions:

- $\|\mathbf{w}_1 - \mathbf{w}^*\| \leq R$ where $\mathbf{w}^* \in \operatorname{argmin}_{\mathbf{w} \in C} \sum_{t=1}^T f_t(\mathbf{w})$.
- $\|\delta f_t(\mathbf{w}_t)\| \leq G$.

■ Theorem: the regret of online projected sub-gradient descent (PSGD) is bounded as follows

$$R_T(\text{PSGD}) \leq \frac{R^2}{2\eta} + \frac{\eta G^2 T}{2}.$$

Choosing η to minimize the bound gives

$$R_T(\text{PSGD}) \leq RG\sqrt{T}.$$

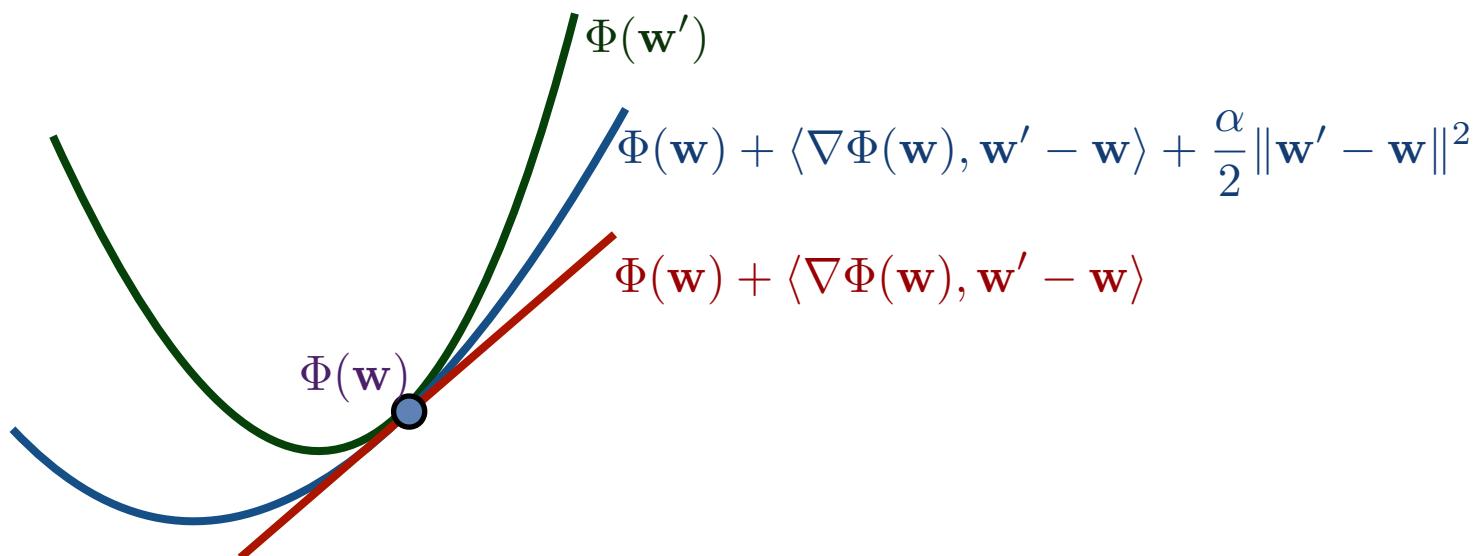
Proof

- The proof uses the definition of subgradient and the property of projection:

$$\begin{aligned} R_T(\text{PSGD}) &= \sum_{t=1}^T (f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*)) \\ &\leq \sum_{t=1}^T \delta f_t(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{w}^*) && \text{(def. of subgrad.)} \\ &= \sum_{t=1}^T \frac{1}{2\eta} \left[\|\mathbf{w}_t - \mathbf{w}^*\|^2 + \eta^2 \|\delta f_t(\mathbf{w}_t)\|^2 - \|\mathbf{w}_t - \eta \delta f_t(\mathbf{w}_t) - \mathbf{w}^*\|^2 \right] \\ &\leq \sum_{t=1}^T \frac{1}{2\eta} \left[\|\mathbf{w}_t - \mathbf{w}^*\|^2 + \eta^2 G^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 \right] && \text{(prop. of proj.)} \\ &\leq \frac{1}{2\eta} \left[\|\mathbf{w}_1 - \mathbf{w}^*\|^2 + \eta^2 G^2 T - \|\mathbf{w}_{T+1} - \mathbf{w}^*\|^2 \right] \\ &\leq \frac{1}{2\eta} \left[\|\mathbf{w}_1 - \mathbf{w}^*\|^2 + \eta^2 G^2 T \right] \leq \frac{1}{2\eta} \left[R^2 + \eta^2 G^2 T \right]. \end{aligned}$$

Strong Convexity

- **Definition:** a convex function Φ defined over a convex set C is α -strongly convex with respect to norm $\|\cdot\|$ if the function $\mathbf{w} \mapsto \Phi(\mathbf{w}) - \frac{\alpha}{2} \|\mathbf{w}\|^2$ is convex or, equivalently,
 - for all \mathbf{w}, \mathbf{w}' in C and $\delta\Phi(\mathbf{w}) \in \partial\Phi(\mathbf{w})$,
$$\Phi(\mathbf{w}') \geq \Phi(\mathbf{w}) + \delta\Phi(\mathbf{w}) \cdot (\mathbf{w}' - \mathbf{w}) + \frac{\alpha}{2} \|\mathbf{w}' - \mathbf{w}\|^2.$$



Strongly Convex Objectives

(Hazan et al., 2007)

- **Theorem:** assume that the functions f_t are α -strongly convex and $\|\delta f_t(\mathbf{w})\| \leq G$ for all \mathbf{w} and $\delta f_t \in \partial f_t(\mathbf{w})$. Then, the regret of online projected sub-gradient descent (PSGD) with parameter $\eta_{t+1} = \frac{1}{\alpha t}$ is bounded as follows

$$R_T(\text{PSGD}) \leq \frac{G^2}{2\alpha} (1 + \log T).$$

Proof

$$R_T(\text{PSGD})$$

$$\begin{aligned}
&= \sum_{t=1}^T (f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*)) \\
&\leq \sum_{t=1}^T \delta f_t(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{w}^*) - \frac{\alpha}{2} \|\mathbf{w}_t - \mathbf{w}^*\|^2 && \text{(strong convexity)} \\
&= \sum_{t=1}^T \frac{1}{2\eta_{t+1}} \left[\|\mathbf{w}_t - \mathbf{w}^*\|^2 + \eta_{t+1}^2 \|\delta f_t(\mathbf{w}_t)\|^2 - \|\mathbf{w}_t - \eta_{t+1} \delta f_t(\mathbf{w}_t) - \mathbf{w}^*\|^2 \right] - \frac{\alpha}{2} \|\mathbf{w}_t - \mathbf{w}^*\|^2 \\
&\leq \sum_{t=1}^T \frac{1}{2\eta_{t+1}} \left[\|\mathbf{w}_t - \mathbf{w}^*\|^2 + \eta_{t+1}^2 G^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 \right] - \frac{\alpha}{2} \|\mathbf{w}_t - \mathbf{w}^*\|^2 && \text{(prop. of proj.)} \\
&\leq \frac{\alpha}{2} \sum_{t=1}^T \left[(t-1) \|\mathbf{w}_t - \mathbf{w}^*\|^2 - t \|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 \right] + \frac{G^2}{2\alpha} \sum_{t=1}^T \frac{1}{t} && \text{(def. of } \eta_{t+1} \text{)} \\
&= \frac{\alpha}{2} \left[-T \|\mathbf{w}_{T+1} - \mathbf{w}^*\|^2 \right] + \frac{G^2}{2\alpha} \sum_{t=1}^T \frac{1}{t} \leq \frac{G^2}{2\alpha} \sum_{t=1}^T \frac{1}{t} \leq \frac{G^2}{2\alpha} (1 + \log T).
\end{aligned}$$

Smoothness

- **Definition:** a continuously differentiable function f is β -smooth if its gradient is β -Lipschitz:

$$\|\nabla f(\mathbf{w}') - \nabla f(\mathbf{w})\| \leq \beta \|\mathbf{w}' - \mathbf{w}\|,$$

for all \mathbf{w}, \mathbf{w}' .

- **Property:** if f is convex and β -smooth, then, for all \mathbf{w}, \mathbf{w}' ,

$$0 \leq f(\mathbf{w}) - f(\mathbf{w}') - \nabla f(\mathbf{w}') \cdot (\mathbf{w} - \mathbf{w}') \leq \frac{\beta}{2} \|\mathbf{w} - \mathbf{w}'\|^2.$$

Exponentiated Gradient (EG)

(Kivinen and Warmuth, 1997)

- Convex set: simplex $C = \{\mathbf{w} \in \mathbb{R}^N : \mathbf{w} \geq 0 \wedge \|\mathbf{w}\|_1 = 1\}$.
- Algorithm:

- $\mathbf{w}_1 = (\frac{1}{N}, \dots, \frac{1}{N})^\top$.
- $\mathbf{w}_{t+1,i} = \frac{\mathbf{w}_{t,i} \exp(-\eta [\delta f_t(\mathbf{w}_t)]_i)}{Z_t}$ where

$$Z_t = \sum_{i=1}^N \mathbf{w}_{t,i} e^{-\eta [\delta f_t(\mathbf{w}_t)]_i}.$$

Analysis

■ Assumption:

- $\|\delta f_t(\mathbf{w}_t)\|_\infty \leq G_\infty$.

■ Theorem: the regret of the Exponentiated Gradient (EG) algorithm is bounded as follows

$$R_T(\text{EG}) \leq \frac{\log N}{\eta} + \frac{\eta G_\infty^2 T}{2}.$$

Choosing η to minimize the bound gives

$$R_T(\text{EG}) \leq 2G_\infty \sqrt{T \log N}.$$

Proof

- Potential: $\Phi_t = D(\mathbf{w}^* \parallel \mathbf{w}_t) = \sum_{i=1}^N \mathbf{w}_i^* \log \frac{\mathbf{w}_i^*}{\mathbf{w}_{t,i}}.$
- $\Phi_{t+1} - \Phi_t = \sum_{i=1}^N \mathbf{w}_i^* \log \frac{\mathbf{w}_{t,i}}{\mathbf{w}_{t+1,i}}$
 $= \sum_{i=1}^N \mathbf{w}_i^* [\log Z_t + \eta[\delta f_t(\mathbf{w}_t)]_i] = \log Z_t + \eta \mathbf{w}^* \cdot \delta f_t(\mathbf{w}_t).$
- $\log Z_t = \log \left[\sum_{i=1}^N \mathbf{w}_{t,i} e^{-\eta[\delta f_t(\mathbf{w}_t)]_i} \right]$
 $= \log \mathbb{E}_{i \sim \mathbf{w}_t} \left[e^{-\eta[\delta f_t(\mathbf{w}_t)]_i} \right]$
 $= \log \mathbb{E}_{i \sim \mathbf{w}_t} \left[e^{-\eta([\delta f_t(\mathbf{w}_t)]_i - \mathbb{E}[[\delta f_t(\mathbf{w}_t)]_i]) - \eta \mathbb{E}[[\delta f_t(\mathbf{w}_t)]_i]} \right]$
 $\leq \eta^2 \frac{4G_\infty^2}{8} - \eta \mathbf{w}_t \cdot \delta f_t(\mathbf{w}_t)$ (Hoeffding's ineq.).

Proof

■ Combining equality and inequality:

$$\begin{aligned}\Phi_{t+1} - \Phi_t &\leq \frac{\eta^2 G_\infty^2}{2} - \eta(\mathbf{w}^* - \mathbf{w}_t) \cdot \delta f_t(\mathbf{w}_t) \\ \Leftrightarrow \eta(\mathbf{w}^* - \mathbf{w}_t) \cdot \delta f_t(\mathbf{w}_t) &\leq \frac{\eta^2 G_\infty^2}{2} + (\Phi_t - \Phi_{t+1}) \\ \Rightarrow \sum_{t=1}^T (\mathbf{w}^* - \mathbf{w}_t) \cdot \delta f_t(\mathbf{w}_t) &\leq \frac{\eta^2 G_\infty^2 T}{2} + \frac{\Phi_1 - \Phi_{T+1}}{\eta} \\ \Rightarrow \sum_{t=1}^T (\mathbf{w}^* - \mathbf{w}_t) \cdot \delta f_t(\mathbf{w}_t) &\leq \frac{\eta^2 G_\infty^2 T}{2} + \frac{\Phi_1}{\eta}. \quad (\text{Rel. Ent. non-neg.})\end{aligned}$$

$$\begin{aligned}R_T(\text{EG}) &= \sum_{t=1}^T (f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*)) \\ &\leq \sum_{t=1}^T \delta f_t(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{w}^*) \\ &\leq \frac{\eta G_\infty^2 T}{2} + \frac{\Phi_1}{\eta} = \frac{\eta G_\infty^2 T}{2} + \frac{D(\mathbf{w}^* \parallel \mathbf{w}_1)}{\eta} \leq \frac{\eta G_\infty^2 T}{2} + \frac{\log N}{\eta}.\end{aligned}$$

Convex Optimization

■ Application: $\min_{\mathbf{w} \in C} f(\mathbf{w})$.

- fixed loss function: $f_t = f$.
- guarantee for average weight vector:

$$\begin{aligned} f\left(\frac{1}{T} \sum_{t=1}^T \mathbf{w}_t\right) - f(\mathbf{w}^*) &\leq \frac{1}{T} \sum_{t=1}^T f(\mathbf{w}_t) - f(\mathbf{w}^*) \\ &= \frac{R_T(\mathcal{A})}{T} = O\left(\frac{1}{\sqrt{T}}\right). \end{aligned}$$

- thus, convergence in $O\left(\frac{1}{\epsilon^2}\right)$.

Generalization

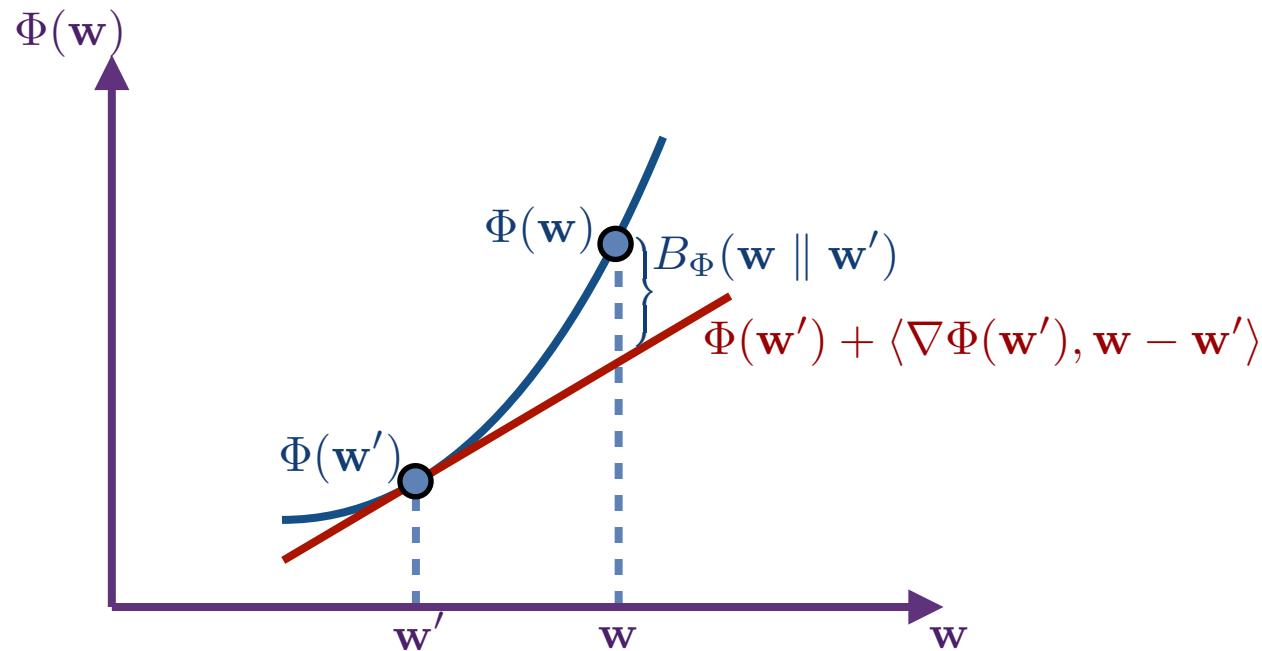
- PSGD and EG both special instances of a more general algorithm: Mirror Descent.
- Mirror Descent is based on Bregman divergence:
 - PSGD: $B(\mathbf{w} \parallel \mathbf{w}') = \frac{1}{2} \|\mathbf{w} - \mathbf{w}'\|_2^2$.
 - EG: unnormalized relative entropy;

$$B(\mathbf{w} \parallel \mathbf{w}') = \sum_{i=1}^N \left[w_i \log \left[\frac{w_i}{w'_i} \right] - w_i + w'_i \right].$$

Bregman Divergence

- **Definition:** Φ convex differentiable over open convex set C .
The Bregman divergence associated to Φ is defined by

$$B_\Phi(\mathbf{w} \parallel \mathbf{w}') = \Phi(\mathbf{w}) - \Phi(\mathbf{w}') - \langle \nabla \Phi(\mathbf{w}'), \mathbf{w} - \mathbf{w}' \rangle.$$



Properties

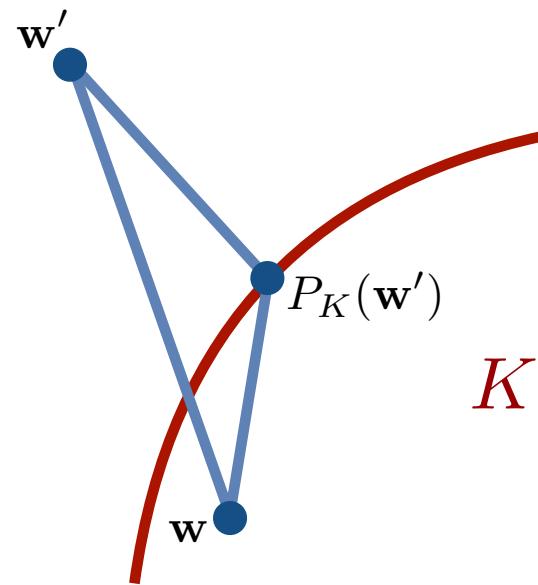
■ **Proposition:** the following properties hold for a Bregman divergence.

- non-negativity: $\forall \mathbf{w}, \mathbf{w}' \in C, B_\Phi(\mathbf{w} \parallel \mathbf{w}') \geq 0$.
- linearity: $B_{\alpha\Phi+\beta\Psi} = \alpha B_\Phi + \beta B_\Psi$.
- projection: for any closed convex set $K \subseteq \overline{C}$, the projection of B_Φ -projection of \mathbf{w}' over K is unique:

$$P_K(\mathbf{w}') = \operatorname{argmin}_{\mathbf{w} \in K} B_F(\mathbf{w} \parallel \mathbf{w}').$$

- Triangular identity:
 $(\nabla\Phi(\mathbf{w}) - \nabla\Phi(\mathbf{v})) \cdot (\mathbf{w} - \mathbf{u}) = B(\mathbf{u} \parallel \mathbf{w}) + B(\mathbf{w} \parallel \mathbf{v}) - B(\mathbf{u} \parallel \mathbf{v})$.
- Pythagorean theorem:
 $B_\Phi(\mathbf{w} \parallel \mathbf{w}') \geq B_\Phi(\mathbf{w} \parallel P_K(\mathbf{w}')) + B_\Phi(P_K(\mathbf{w}') \parallel \mathbf{w}')$.

Pythagorean theorem



$$B_\Phi(\mathbf{w} \parallel \mathbf{w}') \geq B_\Phi(\mathbf{w} \parallel P_K(\mathbf{w}')) + B_\Phi(P_K(\mathbf{w}') \parallel \mathbf{w}').$$

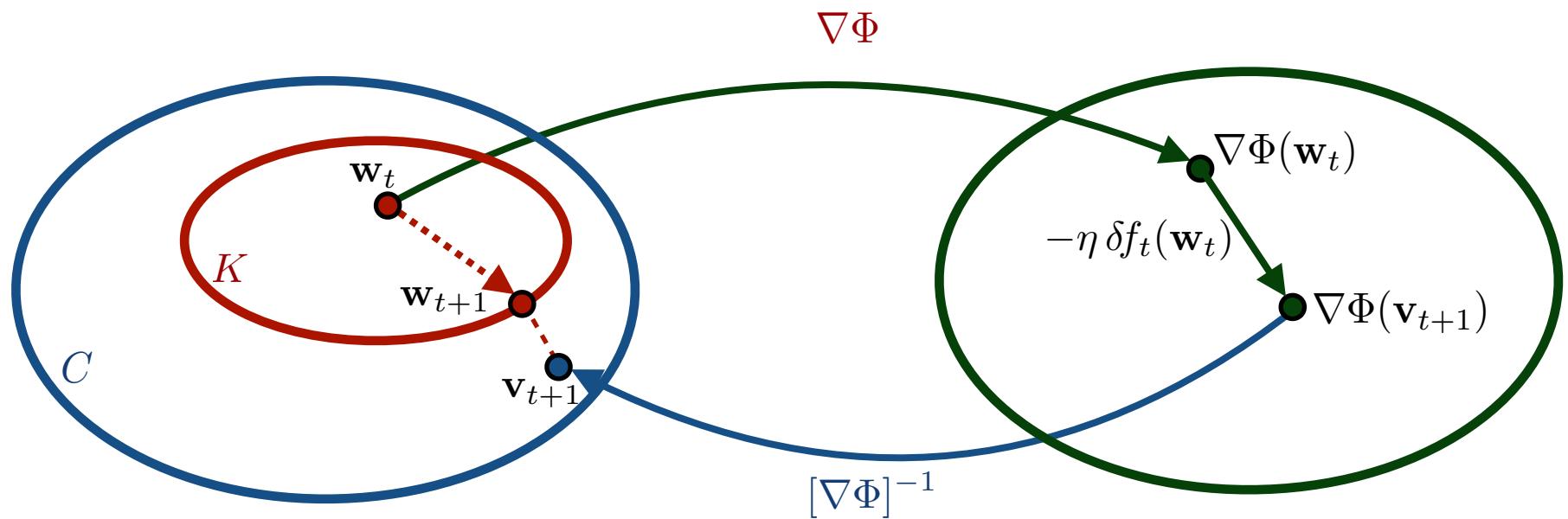
Legendre Type Functions

(Rockafellar, 1970)

- **Definition:** a real-valued function Φ defined over a non-empty open convex set C is said to be of **Legendre type** if it is proper closed convex and differentiable over C and if one of the following equivalent conditions holds:
 - $\nabla\Phi$ is one-to-one mapping from C to $\nabla\Phi(C)$.
 - $\lim_{\mathbf{w} \rightarrow \partial C} \|\nabla\Phi(\mathbf{w})\| = +\infty$.

Mirror Descent

(Nemirovski and Yudin, 1983)



Mirror Descent

MIRROR-DESCENT(Φ)

```
1    $\mathbf{w}_1 \leftarrow \operatorname{argmin}_{\mathbf{w} \in K \cap C} \Phi(\mathbf{w})$ 
2   for  $t \leftarrow 1$  to  $T$  do
3        $\mathbf{v}_{t+1} \leftarrow [\nabla \Phi]^{-1} (\nabla \Phi(\mathbf{w}_t) - \eta \delta f_t(\mathbf{w}_t))$ 
4        $\mathbf{w}_{t+1} \leftarrow \operatorname{argmin}_{\mathbf{w} \in K \cap C} B(\mathbf{w} \parallel \mathbf{v}_{t+1})$ 
```

MD Guarantee

- **Theorem:** let C be a non-empty open convex set and $K \subset \overline{C}$ a compact convex set. Assume that $\Phi: C \rightarrow \mathbb{R}$ is of Legendre type and α -strongly convex with respect to $\|\cdot\|$ and f_t 's convex and G_* -Lipschitz with respect to $\|\cdot\|$. Then, the regret of Mirror Descent can be bounded as follows:

$$R_T(\text{MD}) \leq \frac{B(\mathbf{w}^* \parallel \mathbf{w}_1)}{\eta} + \frac{\eta G_*^2 T}{2\alpha}.$$

Choosing η to minimize the bound gives

$$R_T(\text{MD}) \leq D_\Phi G_* \sqrt{\frac{2T}{\alpha}},$$

with $B(\mathbf{w}^* \parallel \mathbf{w}_1) \leq D_\Phi^2$.

Proof

$$R_T(\text{MD})$$

$$\begin{aligned}
&= \sum_{t=1}^T (f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*)) \\
&\leq \sum_{t=1}^T \delta f_t(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{w}^*) && \text{(def. of subgrad.)} \\
&= \frac{1}{\eta} \sum_{t=1}^T [\nabla \Phi(\mathbf{w}_t) - \nabla \Phi(\mathbf{v}_{t+1})] \cdot (\mathbf{w}_t - \mathbf{w}^*) && \text{(def. of } \mathbf{v}_t) \\
&= \frac{1}{\eta} \sum_{t=1}^T [B(\mathbf{w}^* \parallel \mathbf{w}_t) - B(\mathbf{w}^* \parallel \mathbf{v}_{t+1}) + B(\mathbf{w}_t \parallel \mathbf{v}_{t+1})] && \text{(Breg. div. Identity)} \\
&\leq \frac{1}{\eta} \sum_{t=1}^T [B(\mathbf{w}^* \parallel \mathbf{w}_t) - B(\mathbf{w}^* \parallel \mathbf{w}_{t+1}) - B(\mathbf{w}_{t+1} \parallel \mathbf{v}_{t+1}) + B(\mathbf{w}_t \parallel \mathbf{v}_{t+1})] && \text{(Pythagorean ineq.)} \\
&= \frac{1}{\eta} [B(\mathbf{w}^* \parallel \mathbf{w}_1) - B(\mathbf{w}^* \parallel \mathbf{w}_{T+1})] + \frac{1}{\eta} \sum_{t=1}^T [-B(\mathbf{w}_{t+1} \parallel \mathbf{v}_{t+1}) + B(\mathbf{w}_t \parallel \mathbf{v}_{t+1})] \\
&\leq \frac{B(\mathbf{w}^* \parallel \mathbf{w}_1)}{\eta} + \frac{1}{\eta} \sum_{t=1}^T [B(\mathbf{w}_t \parallel \mathbf{v}_{t+1}) - B(\mathbf{w}_{t+1} \parallel \mathbf{v}_{t+1})].
\end{aligned}$$

Proof

$$\begin{aligned} & \left[B(\mathbf{w}_t \| \mathbf{v}_{t+1}) - B(\mathbf{w}_{t+1} \| \mathbf{v}_{t+1}) \right] \\ &= \Phi(\mathbf{w}_t) - \Phi(\mathbf{w}_{t+1}) - \nabla \Phi(\mathbf{v}_{t+1}) \cdot (\mathbf{w}_t - \mathbf{w}_{t+1}) \\ &\leq (\nabla \Phi(\mathbf{w}_t) - \nabla \Phi(\mathbf{v}_{t+1})) \cdot (\mathbf{w}_t - \mathbf{w}_{t+1}) - \frac{\alpha}{2} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 && (\text{α-strong convexity}) \\ &= -\eta \delta f_t(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{w}_{t+1}) - \frac{\alpha}{2} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 && (\text{def. of } \mathbf{v}_{t+1}) \\ &\leq \eta G_* \|\mathbf{w}_t - \mathbf{w}_{t+1}\| - \frac{\alpha}{2} \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 && (G_*\text{-Lipschitzness}) \\ &\leq \frac{(\eta G_*)^2}{2\alpha}. && (\text{max. of 2nd deg. eq.}) \end{aligned}$$

Equivalent Description

MIRROR-DESCENT(Φ)

```
1    $\mathbf{w}_1 \leftarrow \operatorname{argmin}_{\mathbf{w} \in K \cap C} \Phi(\mathbf{w})$ 
2   for  $t \leftarrow 1$  to  $(T - 1)$  do
3        $\mathbf{w}_{t+1} \leftarrow \operatorname{argmin}_{\mathbf{w} \in K \cap C} \delta f_t(\mathbf{w}_t) \cdot \mathbf{w} + \frac{1}{\eta} B(\mathbf{w} \parallel \mathbf{w}_t)$ 
```

linearization of f_t regularization

■ Proof:

$$\begin{aligned}\mathbf{w}_{t+1} &= \operatorname{argmin}_{\mathbf{w} \in K \cap C} B(\mathbf{w} \parallel \mathbf{v}_{t+1}) \\ &= \operatorname{argmin}_{\mathbf{w} \in K \cap C} \Phi(\mathbf{w}) - \nabla \Phi(\mathbf{v}_{t+1}) \cdot \mathbf{w} \quad (\text{def. of Breg. div.}) \\ &= \operatorname{argmin}_{\mathbf{w} \in K \cap C} \Phi(\mathbf{w}) - (\nabla \Phi(\mathbf{w}_t) - \eta \delta f_t(\mathbf{w}_t)) \cdot \mathbf{w} \quad (\text{def. of } \mathbf{v}_{t+1}) \\ &= \operatorname{argmin}_{\mathbf{w} \in K \cap C} \eta \delta f_t(\mathbf{w}_t) \cdot \mathbf{w} + B(\mathbf{w} \parallel \mathbf{w}_t). \quad (\text{def. of Breg. div.})\end{aligned}$$

Dual Averaging

(Iouditski and Nesterov, 2010)

DUAL-AVERAGING(Φ)

```
1    $\mathbf{v}_1 \leftarrow 0$ 
2    $\mathbf{w}_1 \leftarrow \operatorname{argmin}_{\mathbf{w} \in K \cap C} B(\mathbf{w} \parallel \mathbf{v}_1)$ 
3   for  $t \leftarrow 1$  to  $T$  do
4        $\mathbf{v}_{t+1} \leftarrow [\nabla \Phi]^{-1} (\nabla \Phi(\mathbf{v}_t) - \eta \delta f_t(\mathbf{w}_t))$ 
5        $\mathbf{w}_{t+1} \leftarrow \operatorname{argmin}_{\mathbf{w} \in K \cap C} B(\mathbf{w} \parallel \mathbf{v}_{t+1})$ 
```

Equivalent Description

■ Equivalent form:

$$\begin{aligned}\mathbf{w}_{t+1} &= \operatorname{argmin}_{\mathbf{w} \in K \cap C} B(\mathbf{w} \| \mathbf{v}_{t+1}) \\ &= \operatorname{argmin}_{\mathbf{w} \in K \cap C} \Phi(\mathbf{w}) - \nabla \Phi(\mathbf{v}_{t+1}) \cdot \mathbf{w} \quad (\text{def. of Breg. div.}) \\ &= \operatorname{argmin}_{\mathbf{w} \in K \cap C} \Phi(\mathbf{w}) - \left(\nabla \Phi(\mathbf{v}_t) - \eta \delta f_t(\mathbf{w}_t) \right) \cdot \mathbf{w} \quad (\text{def. of } \mathbf{v}_{t+1}) \\ &= \operatorname{argmin}_{\mathbf{w} \in K \cap C} \eta \sum_{s=1}^t \delta f_t(\mathbf{w}_s) + \Phi(\mathbf{w}). \quad (\text{recurrence})\end{aligned}$$

■ In particular, for linear losses, $f_t(\mathbf{w}) = \mathbf{a}_t \cdot \mathbf{w}$, Dual Averaging coincides with **regularized FL**:

$$\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} \in K \cap C} \sum_{s=1}^t \mathbf{a}_s \cdot \mathbf{w} + \frac{1}{\eta} \Phi(\mathbf{w}).$$

DA Guarantee

- **Theorem:** under the same assumptions as for MD, the following holds for the regret of Dual Averaging,

$$R_T(\text{DA}) \leq \frac{\Phi(\mathbf{w}^*) - \Phi(\mathbf{w}_1)}{\eta} + \frac{2\eta G_*^2 T}{\alpha}.$$

Choosing η to minimize the bound gives

$$R_T(\text{DA}) \leq 2D_\Phi G_* \sqrt{\frac{2T}{\alpha}},$$

with $\Phi(\mathbf{w}^*) - \Phi(\mathbf{w}_1) \leq D_\Phi^2$.

References

- Abraham Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. On-line convex optimization in the bandit setting: gradient descent without a gradient. In SODA, pages 385–394. SIAM, 2005.
- Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *J. Comput. Syst. Sci.*, 74(1):97–114, 2008.
- Amir Beck and Marc Teboulle. Mirror Descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.
- Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.
- A .J. Grove, N. Littlestone, and D. Schuurmans. General convergence results for linear discriminant updates. *Machine Learning*, 43(3):173–210, 2001.

References

- Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.
- Anatoli Iouditski, Yuri Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. 2010. <hal-00508933v1>
- Adam T. Kalai, Santosh Vempala. Efficient algorithms for online decision problems. *J. Comput. Syst. Sci.* 71(3): 291-307. 2005.
- Jyrki Kivinen and Manfred K. Warmuth. Additive versus exponentiated gradient updates for linear prediction. *Information and Computation*, 132(1):1–64, 1997.
- Jyrki Kivinen and Manfred K. Warmuth. Relative loss bounds for multidimensional regression problems. *Machine Learning*, 45(3):301–329, 2001.
- Yurii Nesterov. Introductory lectures on convex optimization: A basic course. Kluwer Academic Publishers, 2004a.
- R. Tyrrell Rockafellar. Convex Analysis. Princeton University Press, 1970.

References

- Arkadii Semenovich Nemirovski, David Berkovich Yudin. Problem complexity and Method Efficiency in Optimization, Wiley, New York, 1983.
- Eiji Takimoto and Manfred K. Warmuth. Path kernels and multiplicative updates. JMLR, 4:773–818, 2003.
- Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In ICML, pages 928–936, 2009.