8 Mirror Descent

Goal

Mismatch between input and gradient spaces, exponentiated gradient, mirror descent, Legendre function.

Alert 8.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.
This note is likely to be updated again soon.

Definition 8.2: Problem

In this lecture we are interested in the constrained minimization problem:

$$\inf_{\mathbf{w}\in C\subseteq \mathsf{V}} \quad f(\mathbf{w}),$$

where f is a convex function that may or may not be smooth. To recall our current theoretical results:

- When f is $\mathsf{L}^{[1]}$-smooth, we obtained the convergence rate $\frac{\mathsf{L}^{[1]}\|\mathbf{w}_0-\mathbf{w}\|_2^2}{2t}$ for the function value in Theorem 4.21.
- When f is $\mathsf{L}^{[0]}$-Lipschitz continuous, we obtained the convergence rate $\frac{\mathsf{L}^{[0]}\|\mathbf{w}_0-\mathbf{w}\|_2}{\sqrt{t}}$ for the (minimum) function value in Theorem 5.17.

The Lipschitz constants and the diameter $\|\mathbf{w}_0-\mathbf{w}\|_2$ both depend on the norm, but there is no reason to believe that the Euclidean norm we used is the best choice. Can we strike a better balance?

Example 8.3: Separable function over simplex

Consider the following simple problem:

$$\min_{\mathbf{w}\in\Delta} \quad \sum_{j=1}^{d} f_j(w_j),$$

where the objective function $f := \sum_j f_j$ is separable in terms of the variables $\mathbf{w} = (w_1, \dots, w_d)$ but the simplex constraint $\Delta = \{\mathbf{w} \in \mathbb{R}^d_+ : \mathbf{1}^\top \mathbf{w} = 1\}$ couples everything. Let us suppose each univariate component function $f_j : \mathbb{R} \to \mathbb{R}$ is 1-Lipschitz continuous. Then, the sum $f : \mathbb{R}^d \to \mathbb{R}$ is $\sqrt{d}$-Lipschitz continuous w.r.t. the Euclidean norm, since

$$\|\nabla f\|_2^2 = \sum_j (\nabla f_j)^2 \le d.$$

The diameter $\|\mathbf{w}_0-\mathbf{w}\|_2 \le \sqrt{2}$. Thus, applying Theorem 5.17 we obtain a convergence rate of $\sqrt{\frac{2d}{t}}$.

Note however that if we choose the norm on $\mathbf{w}$ to be ℓ_1, then

$$\|\nabla f\|_\infty = \max_j |\nabla f_j| \le 1$$

while $\|\mathbf{w}_0-\mathbf{w}\|_1 \le 2$. Can we achieve the convergence rate $\frac{2}{\sqrt{t}}$ by changing the norm? The difference is huge: a factor of square root of the dimension!

Alert 8.4: What makes incremental update possible?

So far, we have seen updates of the following (additive) incremental form:

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \cdot \mathbf{g},$$

which is so natural that we sometimes forget what makes it even mathematically possible:

- the scalar multiplication of the step size η to $\mathbf{g}$;
- the negation $-$;
- and the addition of $\mathbf{w}$ with $-\eta \cdot \mathbf{g}$.

These operations are possible because we have a linear vector space structure in our universe, in particular $\mathbf{w}$ and $\mathbf{g}$ are from the same vector space.

We now make an important distinction: the gradient $\nabla f(\mathbf{w})$ does not come from the same space as $\mathbf{w}$! To be precise, if $\mathbf{w} \in \mathsf{V}$, then the Frechet derivative $f'(\mathbf{w})$ lives in the dual space V^*, i.e., all continuous linear functionals on V. To restore sanity, we need a way to pull things back and forth:

$$\mathsf{J} : \mathsf{V} \to \mathsf{V}^*, \quad \mathsf{J}^{-1} : \mathsf{V}^* \to \mathsf{V}.$$

When we equip the underlying vector space with the Euclidean norm $\|\cdot\|_2$, we may take $\mathsf{J} = \mathsf{J}^* = \mathrm{Id}$, which is the approach we have been taking. In this and the next lecture, we go beyond.

Example 8.5: Exponentiated gradient for online prediction (Kivinen and Warmuth 1997)

Consider forecasting a real quantity $y \in \mathbb{R}$ (e.g. temperature of the day). We consult n experts, each of whom provides a prediction x_i, collectively as $\mathbf{x} \in \mathbb{R}^n$. We then form our own opinion by averaging $\hat{y} = \langle \mathbf{w}, \mathbf{x} \rangle$, $\mathbf{w} \in \Delta$, and suffer the least squares loss $\ell = (y - \hat{y})^2$. Imagine repeating this game for $t = 1, \dots, T$ rounds. What is our average regret compared to the best expert in hindsight?

$$\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2 - \min_{\mathbf{w}\in\Delta} \frac{1}{T}\sum_{t=1}^{T}(y_t - \langle \mathbf{w}, \mathbf{x}_t \rangle)^2, \quad \text{where} \quad \hat{y}_t = \langle \mathbf{w}_t, \mathbf{x}_t \rangle .$$

Surprisingly, Kivinen and Warmuth (1997) showed that the exponentiated gradient (EG) algorithm

$$\tilde{\mathbf{w}}_{t+1} = \mathbf{w}_t \odot \exp(-\eta \ell'(\hat{y}_t - y_t)\mathbf{x}_t)$$
$$\mathbf{w}_{t+1} = \frac{\tilde{\mathbf{w}}_{t+1}}{\langle \mathbf{1}, \tilde{\mathbf{w}}_{t+1} \rangle},$$

achieves diminishing average regret on the order of $O\left(\sqrt{\frac{\ln n}{T}}\right)$, provided that $\|\mathbf{x}_t\|_\infty \le 1$ and $y_t \in [0, 1]$ for all t. To appreciate the significance of this bound, let us note that:

- there is no assumption on how the sequence $\{(\mathbf{x}_t, y_t) : t = 1, \dots, T\}$ is generated! In fact, this sequence can even be adversarial.
- setting $\mathbf{w} = \mathbf{e}_i$ we immediately see that EG performs asymptotically (i.e. when $T \to \infty$) no worse than the best expert *in hindsight*!
- the dependence on the number of experts is only logarithmic! This means we can consult a huge number of experts without deteriorating the bound noticeably.

In contrast, gradient descent achieves a *seemingly* better bound $O(1/\sqrt{T})$, but with the assumption $\|\mathbf{x}_t\|_2 \le 1$ (and a larger pool $\|\mathbf{w}\|_2 \le 1$). In the worst case, $\|\mathbf{x}_t\|_2 \le \sqrt{n}\|\mathbf{x}_t\|_\infty$, leading to the bound $O(\sqrt{n/T})$, which is much worse.

Let us equip $\mathsf{J}(\mathbf{w}) = \ln \mathbf{w}$ (component-wise). Then, we may interpret EG as:

$$\ln \tilde{\mathbf{w}}_{t+1} = \ln \mathbf{w}_t - \eta \ell'(\hat{y}_t - y_t)\mathbf{x}_t,$$

i.e. the usual gradient descent in the (dual) log space.

Kivinen, J. and M. K. Warmuth (1997). "Exponentiated Gradient versus Gradient Descent for Linear Predictors". *Information and Computation*, vol. 132, no. 1, pp. 1–63.

Remark 8.6: Two choices

We now have two choices to address the mismatch between $\mathbf{w} \in \mathsf{V}$ and $\nabla f(\mathbf{w}) \in \mathsf{V}^*$, through a mirror (or duality) map $\mathsf{J} : \mathsf{V} \to \mathsf{V}^*$ with inverse $\mathsf{J}^{-1} : \mathsf{V}^* \to \mathsf{V}$ (hence also the name mirror descent).

- We do our update in the gradient space V^* and pull the update back to the input space V:

$$\mathbf{w}_{t+1} = \mathsf{J}^{-1}(\mathsf{J}(\mathbf{w}_t) - \eta_t \cdot \nabla f(\mathbf{w}_t)). \tag{8.1}$$

  Introducing $\mathbf{w}_t^* := \mathsf{J}(\mathbf{w}_t)$, we can rewrite the above as:

$$\mathbf{w}_{t+1}^* = \mathbf{w}_t^* - \eta_t \cdot \nabla f(\mathsf{J}^{-1}(\mathbf{w}_t^*)).$$

- We pull the gradient back to the input space V and do the update directly there:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \cdot \mathsf{J}^{-1}(\nabla f(\mathbf{w}_t)).$$

We discuss the first approach here and the second one in the next lecture.

Definition 8.7: Legendre function

We call a continuous convex function h Legendre if

- its domain has nonempty interior, i.e., $\operatorname{int}(\operatorname{dom} h) \neq \emptyset$;
- h is differentiable on $\operatorname{int}(\operatorname{dom} h)$;
- $\|\nabla h(\mathbf{w})\| \to \infty$ as $\mathbf{w} \to \partial \operatorname{dom} h$;
- h is strictly convex on $\operatorname{int}(\operatorname{dom} h)$.

Recall that h^* is the Fenchel conjugate of h (see Definition 0.28). It is known that, see e.g. (Bauschke and Borwein 1997),

$$\nabla h : \operatorname{int}(\operatorname{dom} h) \to \operatorname{int}(\operatorname{dom} h^*), \quad \mathbf{w} \mapsto \nabla h(\mathbf{w})$$

is a topological isomorphism, i.e. with continuous inverse $(\nabla h)^{-1} = \nabla h^*$. In other words, we could let $\mathsf{J} = \nabla h$.

Below, we will choose a norm $\|\cdot\|$ and a Legendre function h that is 1-strongly convex w.r.t. $\|\cdot\|$, i.e.

$$\mathsf{D}_h(\mathbf{w}, \mathbf{z}) := h(\mathbf{w}) - h(\mathbf{z}) - \langle \mathbf{w} - \mathbf{z}; \nabla h(\mathbf{z}) \rangle \geq \tfrac{1}{2}\|\mathbf{w} - \mathbf{z}\|^2.$$

Bauschke, H. H. and J. M. Borwein (1997). "Legendre Functions and the Method of Random Bregman Projections". *Journal of Convex Analysis*, vol. 4, no. 1, pp. 27–67.

Example 8.8: (Squared) Euclidean distance is Legendre

Let $h(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$. Verify that h is Legendre and its induced Bregman divergence $\mathsf{D}_h(\mathbf{w}, \mathbf{z}) = \frac{1}{2}\|\mathbf{w} - \mathbf{z}\|_2^2$ is the (square) Euclidean distance. We also have $\mathsf{J}(\mathbf{w}) = \nabla h(\mathbf{w}) = \mathbf{w}$ and of course $\mathsf{J}^{-1} = \mathsf{J}$. This is the choice we have been (implicitly) holding.

Exercise 8.9: KL and Pinsker

Consider the KL function $H(t) = t \ln t - t : \mathbb{R}_+ \to \mathbb{R}$, where $0 \ln 0 := 0$. Verify the following:

- it is indeed Legendre;
- $H'(t) = \ln t$;
- $H^*(s) = \sup_t st - H(t) = \exp(s)$;
- $H' : \mathbb{R}_{++} \to \mathbb{R}$ is continuous and with an continuous inverse;
- define $h(\mathbf{w}) = \sum_j w_j \ln w_j - w_j$. We claim that h, when restricted to the simplex, is 1-strongly convex w.r.t. the ℓ_1 norm. Indeed, we need only verify (see Proposition 6.22):

$$\langle \mathbf{z}; \nabla^2 h(\mathbf{w})\mathbf{z} \rangle = \sum_j z_j^2/w_j \cdot \sum_k w_k \geq (\sum_j |z_j|)^2 = \|\mathbf{z}\|_1^2.$$

The resulting Bregman divergence D_h is known as the KL divergence:

$$\forall \mathbf{w}, \mathbf{z} \geq \mathbf{0}, \quad \mathsf{KL}(\mathbf{w}, \mathbf{z}) = \sum_j w_j \ln \frac{w_j}{z_j} - w_j + z_j,$$

whereas the inequality:

$$\forall \mathbf{w}, \mathbf{z} \in \Delta, \quad \mathsf{KL}(\mathbf{w}, \mathbf{z}) \geq \tfrac{1}{2}\|\mathbf{w} - \mathbf{z}\|_1^2$$

is known as Pinsker's inequality in information theory.

Algorithm 8.10: Mirror descent (MD) (Nemirovski and Yudin 1979)

We now discuss mirror descent using Bregman divergences (see Definition 4.18), an interpretation due to Beck and Teboulle (2003). We define the next iterate as:

$$\begin{aligned}\mathbf{w}_{t+1} &= \operatorname*{argmin}_{\mathbf{w} \in C} \; f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t; \nabla f(\mathbf{w}_t) \rangle + \tfrac{1}{\eta_t} \mathsf{D}_h(\mathbf{w}, \mathbf{w}_t) \geq \; f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t; \nabla f(\mathbf{w}_t) \rangle + \tfrac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|^2 \\ &= \operatorname*{argmin}_{\mathbf{w} \in C} \; \mathsf{D}_h(\mathbf{w}, \mathbf{z}_{t+1}), \quad \text{where} \quad \nabla h(\mathbf{z}_{t+1}) = \nabla h(\mathbf{w}_t) - \eta_t \cdot \nabla f(\mathbf{w}_t),\end{aligned}$$

i.e. the last step projects $\mathbf{z}_{t+1} = (\nabla h)^{-1}(\nabla h(\mathbf{w}_t) - \eta_t \cdot \nabla f(\mathbf{w}_t))$ to the constraint set C using the Bregman divergence D_h (instead of the Euclidean projection). It is clear that the gradient ∇h serves as the mapping J, and the equivalence to (8.1) is apparent (when $C = \mathsf{V}$).

Algorithm: Mirror descent for constrained minimization

Input: $\mathbf{w}_0 \in C$, Legendre function h

for $t = 0, 1, \ldots$ **do**

- compute (sub)gradient $\nabla f(\mathbf{w}_t)$
- choose step size $\eta_t > 0$
- $\nabla h(\mathbf{z}_{t+1}) = \nabla h(\mathbf{w}_t) - \eta_t \cdot \nabla f(\mathbf{w}_t)$ `// update in the gradient space`
- $\mathbf{w}_{t+1} \leftarrow \operatorname*{argmin}_{\mathbf{w} \in C} \mathsf{D}_h(\mathbf{w}, \mathbf{z}_{t+1})$ `// projecting back to the constraint`

If f is $\mathsf{L}^{[1]}$-smooth, then we may choose the step size η_t as in projected gradient Algorithm 3.15 while if f is $\mathsf{L}^{[0]}$-Lipschitz continuous, then we may choose the step size as in the subgradient Algorithm 5.14.

Nemirovski, A. and D. B. Yudin (1979). "Efficient methods for solving large-scale convex programming problems". *Ekonomika i matematicheskie metody*, vol. 15, no. 1, pp. 133–152.
Beck, A. and M. Teboulle (2003). "Mirror descent and nonlinear projected subgradient methods for convex optimization". *Operations Research Letters*, vol. 31, no. 3, pp. 167–175.

Example 8.11: EG belongs to MD

Let $C = \Delta$ and h be the KL function defined in Exercise 8.9. We now compute the Bregman projection:

$$\operatorname*{argmin}_{\mathbf{w}\in\Delta} \mathsf{KL}(\mathbf{w},\mathbf{z}) = \sum_j w_j \log\frac{w_j}{z_j} - w_j + z_j = \sum_j w_j \log\frac{w_j}{z_j/\langle\mathbf{1},\mathbf{z}\rangle} - \log\langle\mathbf{1},\mathbf{z}\rangle - 1 + \langle\mathbf{1},\mathbf{z}\rangle \equiv \mathsf{KL}(\mathbf{w}, \tfrac{\mathbf{z}}{\langle\mathbf{1},\mathbf{z}\rangle}),$$

leading clearly to $\mathbf{w}_+ = \mathbf{z}/\langle\mathbf{1},\mathbf{z}\rangle$. We have already verified that $\nabla h(\mathbf{w}) = \ln\mathbf{w}$ while $(\nabla h)^{-1}(\mathbf{g}) = \exp(\mathbf{g})$ (all component-wise). Thus, the mirror descent step reduces to:

$$\begin{aligned}\mathbf{z}_{t+1} &= (\nabla h)^{-1}(\nabla h(\mathbf{w}_t) - \eta_t\cdot\nabla f(\mathbf{w}_t)) = \mathbf{w}_t\odot\exp(-\eta_t\nabla f(\mathbf{w}_t))\\ \mathbf{w}_{t+1} &= \tfrac{\mathbf{z}_{t+1}}{\langle\mathbf{1},\mathbf{z}_{t+1}\rangle},\end{aligned}$$

which is exactly EG.

The key here is to choose a Legendre function h that matches the "geometry" (i.e. norm) of the constraint set C. Needless to say, there are now infinite possibilities!

Theorem 8.12: Convergence of mirror descent for smooth function

Let $f:\mathbb{R}^d\to\mathbb{R}$ be convex and $\mathsf{L}=\mathsf{L}^{[1]}$-smooth (w.r.t. some norm $\|\cdot\|$), $C\subseteq\mathbb{R}^d$ be closed convex, and η_t is chosen so that (8.2) below holds, then for all $\mathbf{w}\in C$ and $t\geq 1$, the sequence $\{\mathbf{w}_t\}\subseteq C$ generated by Algorithm 8.10 satisfy:

$$f(\mathbf{w}_t)\leq f(\mathbf{w}) + \frac{\mathsf{D}(\mathbf{w},\mathbf{w}_0)}{t\bar\eta_t},\quad \textit{where}\quad \bar\eta_t := \frac{1}{t}\sum_{s=0}^{t-1}\eta_s,$$

where $\mathsf{D}(\mathbf{w},\mathbf{w}_0) = \mathsf{D}_h(\mathbf{w},\mathbf{w}_0)\geq\frac{1}{2}\|\mathbf{w}-\mathbf{w}_0\|^2$ for some 1-strongly convex Legendre function h.

Proof: The proof is literally the same as that of projected gradient. Indeed, using L-smoothness we have for all $\mathbf{w}\in C$:

$$\begin{aligned}f(\mathbf{w}_{t+1}) &\leq f(\mathbf{w}_t) + \langle\mathbf{w}_{t+1}-\mathbf{w}_t;\nabla f(\mathbf{w}_t)\rangle + \tfrac{1}{\eta_t}\mathsf{D}(\mathbf{w}_{t+1},\mathbf{w}_t)\\ &\leq f(\mathbf{w}_t) + \langle\mathbf{w}-\mathbf{w}_t;\nabla f(\mathbf{w}_t)\rangle + \tfrac{1}{\eta_t}\mathsf{D}(\mathbf{w},\mathbf{w}_t) - \tfrac{1}{\eta_t}\mathsf{D}(\mathbf{w},\mathbf{w}_{t+1})\\ &\leq f(\mathbf{w}) + \tfrac{1}{\eta_t}\mathsf{D}(\mathbf{w},\mathbf{w}_t) - \tfrac{1}{\eta_t}\mathsf{D}(\mathbf{w},\mathbf{w}_{t+1}),\end{aligned}\tag{8.2}$$

where the second inequality follows from $\mathbf{w}_{t+1}$ being the Bregman projection to the convex set C, see Proposition 4.20 and Example 4.19, and the last inequality is due to the convexity of f. Take $\mathbf{w}=\mathbf{w}_t$ we see that

$$f(\mathbf{w}_{t+1})\leq f(\mathbf{w}_t),$$

i.e., the algorithm is descending. Summing from $t=0$ to $t=T-1$:

$$T\bar\eta_T\cdot[f(\mathbf{w}_T)-f(\mathbf{w})]\leq\sum_{t=0}^{T-1}\eta_t[f(\mathbf{w}_{t+1})-f(\mathbf{w})]\leq\mathsf{D}(\mathbf{w},\mathbf{w}_0),\quad\text{where}\quad\bar\eta_T := \frac{1}{T}\sum_{t=0}^{T-1}\eta_t,$$

Dividing both sides by $T\bar{\eta}_T$ completes the proof. ■
If there exists a minimizer $\mathbf{w}_\star$, then we have

$$f(\mathbf{w}_t) - f_\star \leq \frac{\mathsf{L}\mathsf{D}(\mathbf{w}_\star, \mathbf{w}_0)}{t}$$

where we have chosen $\eta_t \equiv 1/\mathsf{L}$ to minimize the bound. So the function value converges to the global minimum (thanks to convexity) at the rate of $O(1/t)$. As before, the dependence on L and $\mathbf{w}_0$ makes intuitive sense. Again, the rate of convergence does not depend on d, the dimension!

Theorem 8.13: Convergence of mirror descent for nonsmooth function

Let $C \subseteq \mathbb{R}^d$ be a closed convex set and $f : C \to \mathbb{R}$ be an $\mathsf{L} = \mathsf{L}^{[0]}$-Lipschitz continuous convex function (w.r.t. some norm $\|\cdot\|$). Start with $\mathbf{w}_0 \in C$, for any $\mathbf{w} \in C$, the sequence generated by Algorithm 8.10 satisfies:

$$\min_{0\leq t\leq T-1} f(\mathbf{w}_t) - f(\mathbf{w}) \leq \sum_{t=0}^{T-1} \frac{\eta_t}{\sum_{s=0}^{T-1}\eta_s}(f(\mathbf{w}_t) - f(\mathbf{w})) \leq \frac{2\mathsf{D}(\mathbf{w}, \mathbf{w}_0) + \mathsf{L}^2\sum_{t=0}^{T-1}\eta_t^2}{2\sum_{s=0}^{T-1}\eta_s},$$

where $\mathsf{D}(\mathbf{w}, \mathbf{w}_0) = \mathsf{D}_h(\mathbf{w}, \mathbf{w}_0) \geq \frac{1}{2}\|\mathbf{w} - \mathbf{w}_0\|^2$ for some 1-strongly convex Legendre function h.

Proof: As in the previous proof, since $\mathbf{w}_{t+1}$ is the Bregman projection, we have

$$\begin{aligned}
\langle \mathbf{w}; \nabla f(\mathbf{w}_t)\rangle + \tfrac{1}{\eta_t}\mathsf{D}(\mathbf{w}, \mathbf{w}_t) &\geq \langle \mathbf{w}_{t+1}; \nabla f(\mathbf{w}_t)\rangle + \tfrac{1}{\eta_t}\mathsf{D}(\mathbf{w}_{t+1}, \mathbf{w}_t) + \tfrac{1}{\eta_t}\mathsf{D}(\mathbf{w}, \mathbf{w}_{t+1}) \\
\langle \mathbf{w} - \mathbf{w}_t; \nabla f(\mathbf{w}_t)\rangle + \tfrac{1}{\eta_t}\mathsf{D}(\mathbf{w}, \mathbf{w}_t) &\geq \langle \mathbf{w}_{t+1} - \mathbf{w}_t; \nabla f(\mathbf{w}_t)\rangle + \tfrac{1}{\eta_t}\mathsf{D}(\mathbf{w}_{t+1}, \mathbf{w}_t) + \tfrac{1}{\eta_t}\mathsf{D}(\mathbf{w}, \mathbf{w}_{t+1}) \\
f(\mathbf{w}) - f(\mathbf{w}_t) + \tfrac{1}{\eta_t}\mathsf{D}(\mathbf{w}, \mathbf{w}_t) &\geq -\|\mathbf{w}_{t+1} - \mathbf{w}_t\| \cdot \|\nabla f(\mathbf{w}_t)\|_\circ + \tfrac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 + \tfrac{1}{\eta_t}\mathsf{D}(\mathbf{w}, \mathbf{w}_{t+1}) \\
f(\mathbf{w}) - f(\mathbf{w}_t) + \tfrac{1}{\eta_t}\mathsf{D}(\mathbf{w}, \mathbf{w}_t) &\geq \eta_t\|\nabla f(\mathbf{w}_t)\|_\circ^2/2 + \tfrac{1}{\eta_t}\mathsf{D}(\mathbf{w}, \mathbf{w}_{t+1}).
\end{aligned}$$

Telescoping we obtain

$$\mathsf{D}(\mathbf{w}, \mathbf{w}_T) \leq \mathsf{D}(\mathbf{w}, \mathbf{w}_0) + \sum_{t=0}^{T-1}\eta_t^2\|\nabla f(\mathbf{w}_t)\|_\circ^2/2 + \sum_{t=0}^{T-1}\frac{\eta_t}{\sum_{s=0}^{T-1}\eta_s}(f(\mathbf{w}) - f(\mathbf{w}_t)) \cdot \sum_{s=0}^{T-1}\eta_s.$$

Thus,

$$\min_{0\leq t\leq T-1} f(\mathbf{w}_t) - f(\mathbf{w}) \leq \sum_{t=0}^{T-1} \frac{\eta_t}{\sum_{s=0}^{T-1}\eta_s}(f(\mathbf{w}_t) - f(\mathbf{w})) \leq \frac{2\mathsf{D}(\mathbf{w}, \mathbf{w}_0) + \mathsf{L}^2\sum_{t=0}^{T-1}\eta_t^2}{2\sum_{s=0}^{T-1}\eta_s},$$

as claimed. ■

The bound on the right-hand side vanishes iff $\sum_t \eta_t \to \infty$ and $\eta_t \to 0$.
If we fix a tolerance $\epsilon > 0$ beforehand, then setting $\eta_t = c/\mathsf{L}^2 \cdot \epsilon$ for some constant $c \in]0, 2[$ leads to:

$$\min_{0\leq t\leq T-1} f(\mathbf{w}_t) - f(\mathbf{w}) \leq \epsilon,$$

as long as $T \geq \frac{2\mathsf{L}^2\mathsf{D}(\mathbf{w}, \mathbf{w}_0)}{c(2-c)} \cdot \frac{1}{\epsilon^2}$. The same claim holds for $\bar{\mathbf{w}}_T := \sum_{t=0}^{T-1}\frac{\eta_t}{\sum_{s=0}^{T-1}\eta_s}\mathbf{w}_t$.

Remark 8.14: Composite setting

Duchi and Singer (2009) and Duchi et al. (2010) extended MD to the composite setting where $f = \ell + r$ consists of a smooth component ℓ and a nonsmooth component r, while Duchi et al. (2012) also discussed the stochastic setting.

Duchi, J. C. and Y. Singer (2009). "Efficient Online and Batch Learning Using Forward Backward Splitting". *Journal of Machine Learning Research*, vol. 10, pp. 2899–2934.

Duchi, J. C., S. Shalev-Shwartz, Y. Singer, and A. Tewari (2010). "Composite Objective Mirror Descent". In: *Proceedings of the 23rd Annual Conference on Learning Theory*.

Duchi, J. C., A. Agarwal, M. Johansson, and M. I. Jordan (2012). "Ergodic Mirror Descent". *SIAM Journal on Optimization*, vol. 22, no. 4, pp. 1549–1578.

Remark 8.15: Connection to exponential family and natural gradient

Raskutti and Mukherjee (2015) showed that mirror descent is exactly the natural gradient algorithm on a dual Riemannian manifold, which is expected given the gradient space update interpretation of mirror descent. Kunstner et al. (2021) also connected mirror descent with EM for exponential families.

Raskutti, G. and S. Mukherjee (2015). "The Information Geometry of Mirror Descent". *IEEE Transactions on Information Theory*, vol. 61, no. 3, pp. 1451–1457.

Kunstner, F., R. Kumar, and M. Schmidt (2021). "Homeomorphic-Invariance of EM: Non-Asymptotic Convergence in KL Divergence for Exponential Families via Mirror Descent". In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 3295–3303.