CPSC 536C: Algorithms for Convex Optimization
University of British Columbia

Mirror Descent

Akshay Ram
March 9, 2025

These notes are subject to change and may contain errors.

1 Introduction

In the previous note we studied the simplest first-order method: gradient descent. The analysis of this method depended on certain assumptions about the convex function: e.g. Lipschitz condition or strong convexity or smoothness. In this note we will generalize both the algorithm and analysis to the setting of different norms. This will be useful in giving a formal derivation of the Multiplicative Weights method as a natural first-order method.

1.1 Bregman Divergence

Recall our intuition for gradient descent update

$$x_{t+1} = \arg\min_{x \in K} f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta}\|x - x_t\|_2^2.$$

The first two terms are a linear approximation for f at our previous iterate x_t; the final 'regularizer' term is used to reflect that the function is only close to linear near x_t, so we have less knowledge about the value of the function as we get further from x_t.

In Mirror Descent we consider what happens if we use a different regularizer in order to incorporate different geometry.

Definition 1. *Let $\phi : dom(\phi) \to \mathbb{R}$ be a differentiable convex function, then the Bregman divergence associated to ϕ is*

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

This is just the difference between the value $\phi(y)$ and the value given by the linear approximation at x. The following facts are straightforward:

Fact 2. *For differentiable convex $\phi : dom(\phi) \to \mathbb{R}$, the Bregman divergence D_ϕ satisfies*

1. *$D_\phi(y \mid x) \geq 0$ for all $x, y \in dom(\phi)$;*
2. *$\nabla_y D_\phi(y \mid x) = \nabla\phi(y) - \nabla\phi(x)$;*
3. *$D_\phi(y \mid x)$ is a convex function of y (but not necessarily of x).*
4. *Let $r(y) := D_\phi(y \mid x)$; then $D_r(z \mid w) = D_\phi(z \mid w)$.*

The following natural examples show that Bregman divergence will allow us to generalize the analysis of gradient descent from the Euclidean setting to other notions of distance.

Exercise 1. *For the following functions ϕ, verify they are differentiable and convex, and compute their Bregman divergence D_ϕ:*

- *Affine $\phi(x) := \langle b, x\rangle + c$;*
- *Quadratic $\phi(x) := \langle x, Qx\rangle + \langle b, x\rangle + c$ where $Q \succeq 0$;*
- *Check that for $\phi(x) = \frac{1}{2}\|x\|_2^2$, $D_\phi(y \mid x) = \frac{1}{2}\|y - x\|_2^2$;*
- *For $1 < p \leq 2 : \phi(x) := \frac{1}{2}\|x\|_p^2$;*
- *For $q \geq 2 : \phi(x) := \frac{q}{q-1}\sum_i x_i^{1-1/q}$;*
- *Entropy $\phi(p) := \sum_i p_i \log p_i$ for $p \geq 0$.*

1.2 Assumptions

In this note we will generalize the previous assumptions to arbitrary norms, which will allow us to use Bregman divergences to generalize our previous analysis to Mirror Descent.

Definition 3. *$f : \mathbb{R}^n \to \mathbb{R}$ is G-Lipschitz wrt norm $\|\cdot\|$ if*

$$\forall x, y : \quad |f(y) - f(x)| \leq G\|y - x\|.$$

Note that this is equivalent to $\|g\|_* \leq G$ for all x and $g \in \partial f(x)$, where $\|g\|_* := \sup_{\|z\|\leq 1}\langle g, z\rangle$ is the dual norm.

This intuitively allows us to relate f to an affine function with slope G. We can also relate f to quadratic functions of different norms.

Definition 4. *$f : \mathbb{R}^n \to \mathbb{R}$ is α-strongly convex wrt $\|\cdot\|$ if*

$$\forall x, y : \quad f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|y - x\|^2.$$

f is β-smooth wrt $\|\cdot\|$ if

$$\forall x, y : \quad f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\beta}{2}\|y - x\|^2.$$

In both cases this allows us to compare our function to $\|\cdot\|^2$ which is a simpler quadratic and should be easier to analyze. The point of the above assumptions are that they allow us to relate various notions of approximate solutions.

Fact 5. *Let f be a convex function which is (1) G-Lipschitz, (2) β-smooth, and (3) α-strongly convex, respectively, all wrt norm $\|\cdot\|$, and let x^* be the optimizer of f. Also let ϕ be a differentiable convex function that is 1-strongly convex wrt $\|\cdot\|$. Then*

1. *$|f(y) - f(x)| \leq G\sqrt{D_\phi(y \mid x)}$;*
2. *$f(x) - f(x^*) \leq \beta D_\phi(x \mid x^*)$;*

3. *$f(x) - f(x^*) \leq \frac{\|\nabla f(x)\|_*^2}{2\alpha}$ and $f(x) - f(x^*) \geq \frac{\alpha}{2}\|x - x^*\|^2$.*

Proof:

1. Follows directly from the definition;

2. Apply the definition of smoothness between x and x^*:

$$f(x) \leq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \frac{\beta}{2}\|x - x^*\|^2 = f(x^*) + \frac{\beta}{2}\|x - x^*\|^2,$$

   where in the last step we used $\nabla f(x^*) = 0$ since x^* is the optimizer.

3. Rearranging, we see that this is equivalent to a lower bound on $f(x^*)$ in terms of the gradient norm. For this we use our strong convexity assumption

$$\begin{aligned} f(x^*) &\geq f(x) + \langle \nabla f(x), x^* - x \rangle + \frac{\alpha}{2}\|x^* - x\|^2 \\ &\geq f(x) - \|\nabla f(x)\|_* \|x^* - x\| + \frac{\alpha}{2}\|x^* - x\|^2 =: q_\alpha(\|x^* - x\|). \end{aligned}$$

   Then we have

$$f(x^*) q_\alpha(\|x^* - x\|) \geq \min_r q_\alpha(r) = f(x) - \frac{\|\nabla f(x)\|_*^2}{2\alpha}.$$

   The result follows by rearranging.

   For the lower bound, we can simply apply the definition

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \frac{\alpha}{2}\|x - x^*\|^2 = f(x^*) + \frac{\alpha}{2}\|x - x^*\|^2$$

   as $\nabla f(x^*) = 0$ at the optimizer.

□

Exercise 2.

- *Differentiable convex ϕ is α-strongly convex and β-smooth wrt norm $\|\cdot\|$ iff*

$$\frac{\alpha}{2}\|y - x\|^2 \leq D_\phi(y \mid x) \leq \frac{\beta}{2}\|y - x\|^2.$$

- *For $\phi(x) = \frac{1}{2}\|x\|_2^2$, the Bregman divergence $D_\phi(y \mid x) = \frac{1}{2}\|y - x\|_2^2$ is 1-strongly convex and 1-smooth wrt the Euclidean norm $\|\cdot\|_2$.*

- *(Pinsker's Inequality: difficult exercise) For negative entropy $\phi(p) := \sum_i p_i \log p_i$, the Bregman divergence is the Kullback-Leibler divergence, or relative entropy $D_\phi(q \mid p) = \sum_i q_i \log \frac{q_i}{p_i}$. If p, q are probability distributions (i.e. $p, q \geq 0, \langle p, 1_n \rangle = \langle q, 1_n \rangle = 1$, then*

$$D_\phi(q \mid p) \geq \frac{1}{2}\|q - p\|_1^2.$$

  This implies ϕ is 1-strongly convex wrt $\|\cdot\|_1$.

2 Mirror Descent

The algorithm in this section will use a Bregman divergence in place of the Euclidean norm squared in order to compute the update:

$$\tilde{x}_{t+1} := \arg\min_x f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{\eta} D_\phi(x \mid x_t); \tag{1}$$

$$x_{t+1} := \arg\min_{x \in K} D(x \mid \tilde{x}_{t+1}). \tag{2}$$

Fact 6. *In the above setting, under some technical conditions*

$$\nabla\phi(\tilde{x}_{t+1}) = \nabla\phi(x_t) - \eta \nabla f(x_t).$$

This can be verified directly using the optimality conditions. The technical conditions are needed to make sure such a solution exists, i.e. that there is an element x such that $\nabla\phi(x) = \nabla\phi(x_t) - \eta\nabla f(x_t)$. Note that the set of gradients $\{\nabla\phi(x)\} = \text{dom}(\phi^*)$, so this for example gives a condition on the step-size η so that the gradients remain within this domain. The above fact is also why we can think of mirror descent as a 'dual' version of gradient descent: we are using the gradient of f to update our iterate in the dual space of gradients of ϕ, and then using our 'mirror map' $(\nabla\phi)^{-1}$ to induce our primal iterates x to match these dual ones.

In some cases, like gradient descent with the Euclidean norm, this two-step update can be re-interpreted as a single step constrained update. We will revisit this in specific applications.

In order to analyze the algorithm, we can re-interpret this two-step update in terms of a constrained version of the quadratic proxy update used in gradient descent:

Claim 7. *If K has a relative interior, then the above update can be equivalently written as*

$$y_{t+1} = \arg\min_{x \in K} q(x) := f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{\eta} D(x \mid x_t).$$

Proof: We can verify that this update satisfies the KKT conditions for the two-step update, which by Slater's condition is equivalent to optimality.

$$y^* = \arg\min_{y \in K} f(y) \iff \quad 0 \in \partial\delta_K(y^*) + \nabla f(y^*) = 0.$$

So we first compute the optimality condition of the two-step update above:

$$\tilde{x}_{t+1} = \arg\min_x q(x) \iff \nabla f(x_t) + \frac{\nabla\phi(\tilde{x}_{t+1}) - \nabla\phi(x_t)}{\eta} = 0 \iff \nabla\phi(\tilde{x}_{t+1}) = \nabla\phi(x_t) - \eta\nabla f(x_t),$$

$$x_{t+1} = \arg\min_{x \in K} D_\phi(x \mid \tilde{x}_{t+1}) \iff \exists g \in \partial\delta_K(x_{t+1}), \nabla\phi(x_{t+1}) - \nabla\phi(\tilde{x}_{t+1}) = -g,$$

where the first line was by simple optimality for convex q with $\nabla_y D(y \mid x) = \nabla\phi(y) - \nabla\phi(x)$ as given in theorem 2, and the second line was by KKT optimality conditions.

Next we compute the optimality condition for the single step update:

$$y_{t+1} = \arg\min_{x \in K} q(x) \iff \nabla q(y_{t+1}) = \nabla f(x_t) + \frac{\nabla\phi(y_{t+1}) - \nabla\phi(x_t)}{\eta} \in \partial\delta_K(y_{t+1}).$$

But we note that $\nabla\phi(\tilde{x}_{t+1}) = \nabla\phi(x_t) - \eta\nabla f(x_t)$, and that the sub-differential $\partial\delta_K$ for an indicator function is a cone ($g \in \partial\delta_K(y) \iff \lambda g \in \partial\delta_K(y) \forall \lambda > 0$); so we can rearrange the above

$$y_{t+1} = \arg\min_{x \in K} q(x) \iff \nabla f(x_t) + \nabla\phi(y_{t+1}) - \nabla\phi(\tilde{x}_{t+1}) \in \partial\delta_K(y_{t+1}).$$

But then the KKT certificate (x_{t+1}, g) from the two-step update also gives a KKT certificate for this update step, verifying optimality $y_{t+1} = x_{t+1}$.

Now we can perform essentially the same analysis as gradient descent. We emphasize here that the algorithm requires computing Bregman projection $\arg\min_{x \in K} D(x \mid x_t)$, which in itself may be computationally difficult. But if this can be accomplished, then the convergence guarantee is the same as the unconstrained setting.

In order to analyze the two-step update, the following lemma is super useful.

Proposition 8 (Prop 2.20 in Yaoliang's notes)**.** *Let $q = \ell + r$ where ℓ, r are both convex and r is differentiable. Then*

$$x_* \in \arg\min_x q(x) \quad \iff \quad \forall x \in \mathbb{R}^d : q(x) \geq q(x_*) + D_r(x \mid x_*).$$

Proof: Clearly if $q(x) \geq q(x_*) + D_r(x \mid x_*)$ for all x then $x_* \in \arg\min F$ by non-negativity of Bregman divergence D_r. Conversely, if $x_* \in \arg\min_x q(x)$ then 0 is a subgradient at x^*, i.e.

$$0 \in \partial q(x_*) = \partial \ell(x^*) + \nabla r(x^*),$$

where we used convexity of q and differentiability of r. Rearranging, this shows $-\nabla r(x^*) \in \partial \ell(x^*)$, so we can lower bound

$$\forall x \in \mathbb{R}^d : \ell(x) \geq \ell(x_*) - \langle \nabla r(x_*), x - x_* \rangle.$$

Putting this together gives

$$\begin{aligned} q(x) &= \ell(x) + r(x) \geq \ell(x_*) - \langle \nabla r(x^*), x - x^* \rangle + r(x) \\ &= \ell(x^*) + r(x^*) + (r(x) - r(x^*)) - \langle \nabla r(x^*), x - x^* \rangle = q(x^*) + D_r(x \mid x^*), \end{aligned}$$

where the second step was by the subgradient bound above for ℓ, and in the last step we used the definition of $q = \ell + r$ and Bregman divergence D_r. □

The following corollary is used to analyze the projection step and is left as an exercise.

Corollary 9. *Let $x^* = \arg\min_{y \in K} D_\phi(y \mid x)$. Then for any $y \in K$ we have*

$$D_\phi(y \mid x) \geq D_\phi(x^* \mid x) + D_\phi(y \mid x^*).$$

3 Analysis of Mirror Descent

We first show the analysis for the Lipschitz setting.

Theorem 10. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and G-Lipschitz wrt norm $\|\cdot\|$, let $K \subseteq \mathbb{R}^n$ be a convex compact constraint set. Given regularizer ϕ that is 1-strongly convex wrt $\|\cdot\|$, consider initial guess x_0 and update step*

$$\tilde{x}_{t+1} := \arg\min_x f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{\eta} D_\phi(x \mid x_t); \qquad x_{t+1} := \arg\min_{x \in K} D(x \mid \tilde{x}_{t+1}).$$

Then with $\eta \approx \sqrt{\frac{D(x^* | x_0)}{G^2 T}}$, Mirror descent achieves convergence

$$\min_{t \in [T]} f(x_t) - \min_{x \in K} f(x) \lesssim G \sqrt{\frac{D_\phi(x^* \mid x_0)}{T}}.$$

Proof: We begin in the same way as gradient descent analysis, replacing the Euclidean norm regularizer with a Bregman projection. So for any $x \in K$:

$$\begin{aligned}
f(x) + \frac{1}{\eta} D(x \mid x_t) &\geq f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{\eta} D(x \mid x_t) \\
&\geq f(x_t) + \langle \nabla f(x_t), \tilde{x}_{t+1} - x_t \rangle + \frac{1}{\eta} D(\tilde{x}_{t+1} \mid x_t) + \frac{1}{\eta} D(x \mid \tilde{x}_{t+1}) \\
&\geq f(x_t) + \langle \nabla f(x_t), \tilde{x}_{t+1} - x_t \rangle + \frac{1}{\eta} D(\tilde{x}_{t+1} \mid x_t) + \frac{1}{\eta} (D(x_{t+1} \mid \tilde{x}_{t+1}) + D(x \mid x_{t+1})) \\
&\geq f(x_t) - \|\nabla f(x_t)\|_* \|\tilde{x}_{t+1} - x_t\| + \frac{1}{2\eta} \|\tilde{x}_{t+1} - x_t\|^2 + \frac{1}{\eta} D(x \mid x_{t+1}) \\
&\geq f(x_t) - \frac{\eta \|\nabla f(x_t)\|_*^2}{2} + \frac{1}{\eta} D(x \mid x_{t+1}),
\end{aligned}$$

where the first step is by convexity, in the second we applied theorem 8 with $r(x) := \eta^{-1} D(x \mid x_t)$ (noting $D_r = D_\phi$ by theorem 2(4)) and $\ell(x) := f(x_t) + \langle \nabla f(x_t), x - x_t \rangle$ for which the optimizer is $\tilde{x}_{t+1} = \arg\min_x q(x)$, in the third we applied theorem 8 or specifically theorem 9 to lower bound using the projection step, in the fourth step we applied duality of norms for the inner product term and strong convexity of ϕ for the Bregman divergence term as well as $D(x_{t+1} \mid \tilde{x}_{t+1}) \geq 0$, and in the last step we lower bounded by the minimum of the quadratic function in $\|\tilde{x}_{t+1} - x_t\|$. Rearranging this step gives

$$f(x_t) - f(x) \leq \frac{\eta \|\nabla f(x_t)\|_*^2}{2} + \frac{1}{\eta} (D(x \mid x_t) - D(x \mid x_{t+1})).$$

In order to bound the best iterate we apply the above at $x = x^*$ and consider the average

$$\begin{aligned}
\frac{1}{T} \sum_{t=0}^{T-1} f(x_t) - f(x^*) &\leq \frac{1}{T} \sum_{t=0}^{T-1} \frac{\eta \|\nabla f(x_t)\|_*^2}{2} + \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{\eta} (D(x^* \mid x_t) - D(x^* \mid x_{t+1})) \\
&\leq \frac{\eta G^2}{2} + \frac{D(x^* \mid x_0) - D(x^* \mid x_T)}{\eta T} \lesssim \sqrt{\frac{G^2 D(x^* \mid x_0)}{T}},
\end{aligned}$$

where the first step was by the bound above, in the second step we used G-Lipschitzness for the first term and telescoping for the second, and in the final step we used $D(x^* \mid x_T) \geq 0$ and set $\eta \approx \sqrt{\frac{D(x^*|x_0)}{G^2 T}}$ to balance terms. We can now bound the best iterate by the average to give the stated convergence result. □

Next we consider the smooth setting.

Theorem 11. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and β-smooth wrt norm $\|\cdot\|$. Given regularizer ϕ that is 1-strongly convex wrt $\|\cdot\|$, consider initial guess x_0 and update step*

$$x_{t+1} := \arg\min_x D(x \mid x_t)$$

Then Mirror descent achieves convergence

$$f(x_T) - \min_{x \in K} f(x) \lesssim \frac{\beta D_\phi(x^* \mid x_0)}{T}.$$

Proof: We follow the same steps but use the smoothness guarantee to choose the parameter for the regularizer:

$$\begin{aligned} f(x) + \beta D(x \mid x_t) &\geq f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \beta D(x \mid x_t) \\ &\geq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \beta D(x_{t+1} \mid x_t) + \beta D(x \mid x_{t+1}) \\ &\geq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{\beta}{2}\|x_{t+1} - x_t\|^2 + \beta D(x \mid x_{t+1}) \\ &\geq f(x_{t+1}) + \beta D(x \mid x_{t+1}), \end{aligned}$$

where the first step is by convexity, in the second we applied theorem 8 with $r(x) := \beta D(x \mid x_t)$ (noting that $D_r = D_\phi$ by theorem 2(4)) and $\ell(x) := f(x_t) + \langle \nabla f(x_t), x - x_t \rangle$ with optimizer $x_{t+1} = \arg\min_x q(x)$, in the third step we applied 1-strong convexity of ϕ, which then gives the last step by our smoothness assumption on f. Rearranging gives

$$f(x_{t+1}) - f(x) \leq \beta(D(x \mid x_t) - D(x \mid x_{t+1})).$$

Applying this with $x = x^*$ and considering the average gives

$$\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) \leq \frac{1}{T}\sum_{t=1}^{T} \beta(D(x^* \mid x_{t-1}) - D(x^* \mid x_t)) \leq \frac{\beta D(x^* \mid x_0)}{T},$$

where the first step was by the bound above, in the second step we applied telescoping, and in the final step we used $D(x^* \mid x_T) \geq 0$. Finally, apply the above bound with $x = x_t$ gives $f(x_{t+1}) \leq f(x_t)$, so the average bound gives the last iterate bound directly. □

4 Online Optimization

We now take a detour to a more general problem: online optimization. This perspective will help us with applications given at the end of the note. The formal problem setting is as follows:

Definition 12. *At each time step the algorithm chooses an action $x_t \in \mathcal{X}$. Subsequently the adversary (or nature) produces a loss function $\ell_t : \mathcal{X} \to \mathbb{R}$. The goal is to minimize our* regret

$$R(x_1, ..., x_T) := \sum_{t \in [T]} \ell_t(x_t) - \min_{x \in \mathcal{X}} \sum_{t \in [T]} \ell_t(x),$$

where we are comparing to the best single action in hindsight.

This is a very general framework with many applications. An important special case is the setting of online prediction with experts. Here there are n experts, the action space corresponds to probability distributions over the experts, and the feedback loss is a linear function over the experts. In this setting the most famous algorithm is known as Multiplicative Weights:

Definition 13. *Let $\ell_t(p) := \langle g^t, p \rangle$ be linear loss functions satisfying $g^t \in [-a, b]^n$ where $b \geq a \geq 0$ without loss. The MW algorithm with learning rate η maintains starting weights $w^0 := 1_n$ and distribution $p^0 := 1_n/n$, and in each iteration updates*

$$\frac{w_i^{t+1}}{w_i^t} := \begin{cases} (1-\eta)^{g_i^t/b} & g_i^t \geq 0 \\ (1+\eta)^{-g_i^t/b} & g_i^t \leq 0 \end{cases}, \qquad \Phi^{t+1} := \langle w^{t+1}, 1_n \rangle, \quad p^{t+1} := \frac{w^{t+1}}{\Phi^{t+1}}.$$

Theorem 14 (Arora, Hazan, Kale Theorem 2)**.** *For learning rate $\eta \leq 1/2$, the algorithm produces regret bound*

$$\forall i \in [n]: \quad \sum_t \ell_t(p^t) - \ell_t(e_i) \leq \eta \sum_{t \in [T]} |\ell_t(e_i)| + \frac{b \log n}{\eta}.$$

If $g^t \in [0, b]^n$, then we get multiplicative bound

$$\forall i \in [n]: \quad \sum_t \ell_t(p^t) \leq (1+\eta) \sum_t \ell_t(e_i) + \frac{b \log n}{\eta}.$$

Proof: The first statement if proved in the survey of Arora, Hazan, Kale. We repeat the proof of the second since it gives a slightly more refined bound for multiplicative error.

The analysis proceeds by using Φ^t as a potential function. First we can upper bound the change

$$\frac{\Phi^{t+1}}{\Phi^t} = \sum_i \frac{w_i^t (1-\eta)^{g_i^t/b}}{\Phi^t} = \sum_i p_i^t (1-\eta)^{g_i^t/b} \leq \langle p^t, 1_n - \eta g^t/b \rangle \leq \exp(-\eta \langle p^t, g^t \rangle / b),$$

where the first step was by definition of Φ^t, w^t, in the third step we used Taylor approximation $(1-\eta)^x \leq 1 - \eta x$ for $\eta \leq 1/2, x \in [0,1]$, and in the final step we used convexity to bound $1 - x \leq e^{-x}$.

At the end of the procedure, we can lower bound the potential based on the best expert

$$\Phi^T = \sum_i w_i^T \geq w_i^T = \exp(\sum_t g_i^t / b \log(1-\eta)),$$

where the last step was again by definition of the updates. Finally, putting these together gives

$$\sum_t g_i^t / b(-\eta - \eta^2) \leq \log w_t^T \leq \log \Phi^T \leq \log n - \eta \sum_t \langle p^t, g^t/b \rangle$$

$$\implies \sum_t \langle p^t, g^t \rangle \leq (1+\eta) \sum_t \langle e_i, g^t \rangle + \frac{b \log n}{\eta},$$

where in the first line we used Taylor approximation $\log(1-\eta) \geq -\eta - \eta^2$ for $\eta \leq 1/2$, and the last line was by rearranging. □

The particular analysis above can be used to give quite a refined bound for particular structured loss functions. Exponential Hedge is a slightly modified version of Multiplicative Weights that achieves similar regret.

Definition 15. *Let $\ell_t(p) := \langle g^t, p \rangle$ be linear loss functions. The Exponential Hedge algorithm with learning rate η maintains starting weights $w^0 := 1_n$ and distribution $p^0 := 1_n/n$, and in each iteration updates*

$$\frac{w_i^{t+1}}{w_i^t} := \exp(-\eta g_i^t), \qquad \Phi^{t+1} := \langle w^{t+1}, 1_n \rangle, \quad p^{t+1} := \frac{w^{t+1}}{\Phi^{t+1}}.$$

A simple but very powerful observation is that the Mirror Descent algorithm can be performed in an online fashion, and the analyses the Lipschitz setting in fact produce equivalent regret bounds for online convex optimization. We give the formal statement below without proof:

Theorem 16. *In the online optimization setting let* $\{\ell_t(x) := \langle g_t, x\rangle\}$ *be linear loss functions; further let* K *be the constraint set and assume access to regularizer* ϕ *that is* 1*-strongly convex wrt* $\|\cdot\|$*. Then the Mirror descent algorithm above gives regret bound wrt any* $x \in K$

$$\sum_{t\in[T]} \langle g_t, x_t - x\rangle \le \sum_{t\in[T]} \frac{\eta\|g_t\|_*^2}{2} + D(x \mid x_0).$$

Further, if $g_t \in \partial f_t(x_t)$ *for convex functions* $\{f_t\}$*, then we get regret bound*

$$\sum_{t\in[T]} f_t(x_t) - f_t(x) \le \sum_{t\in[T]} \langle g_t, x_t - x\rangle \le \sum_{t\in[T]} \frac{\eta\|g_t\|_*^2}{2} + D(x \mid x_0).$$

It turns out that the Hedge algorithm is a special case of online mirror descent using Entropy regularizer.

Theorem 17 (Expert Prediction with Hedge)**.** *The* Exponential Hedge *algorithm with learning rate* η *produces distributions* $p_1, ..., p_T \in \Delta_n$ *with*

$$\sum_{t\in[T]} \langle p_t, g_t\rangle - \min_{p\in\Delta} \sum_{t\in[T]} \langle p, g_t\rangle \le O(\eta) \sum_{t\in[T]} \|g_t\|_\infty^2 + \frac{\log n}{\eta}.$$

Proof: We use Bregman divergence for entropy regularizer

$$\phi(p) := \sum_i p_i \log p_i, \qquad D_\phi(q \mid p) = \sum_i q_i \log \frac{q_i}{p_i}.$$

We can verify that the update steps have the following explicit form

$$\tilde{p}_i^{t+1} = p_i^t \exp(-\eta g_i^t), \qquad p^{t+1} = \frac{\tilde{p}^{t+1}}{\langle \tilde{p}^{t+1}, 1_n\rangle}.$$

Further, Pinsker's inequality gives that ϕ is 1-strongly convex wrt $\|\cdot\|_1$, so we can use Lipschitz condition $\|g\|_* = \|g\|_\infty$. Plugging this into the Mirror descent guarantee gives $\forall p \in \Delta$:

$$\sum_{t\in[T]} \langle g^t, p^t - p\rangle \le O(\eta) \sum_{t\in[T]} \|g_t\|_\infty^2 + \frac{D(p \mid p^0)}{\eta},$$

and the result follows by noting $D(p \mid p^0) \le \log n$ for uniform distribution $p^0 := 1_n/n$. □

The magic of the above algorithms is that they perform well compared to the offline optimum, even though the adversary gets to choose feedback after learning the action of the algorithm. Rearranging the bound, this tells us that if we play as the adversary, we can choose feedback to guarantee that each online step is good, and the online algorithm will guarantee that this implies the eventual offline solution is also good. This will be useful for our applications.

5 Applications

The above perspective will allow us to use Mirror descent as a dual optimization algorithm.

5.1 LP Solver

We consider the simple feasibility problem: find $x \in K$ such that $Ax \leq b$. Here K should be thought of as 'easy' constraints (such as non-negativity or ball constraints), and $Ax \leq b$ will be more difficult to deal with. We first reframe this as an optimization problem:

$$\exists x \in K : Ax \leq b \iff \min_{x \in K} f(x) := \max_i \langle a_i, x \rangle - b_i = \max_{p \in \Delta} \langle p, Ax - b \rangle \leq 0.$$

So we can instead consider optimizing the convex function f given above. But this function is non-differentiable and incorporates all constraints simultaneously, which may be difficult to deal with. Therefore, we use Mirror descent, and specifically multiplicative weights or the Exp Hedge algorithm, to update our dual iterate $p \in \Delta$ as a linear proxy to the true function f. Note that $f(x)$ is exactly the offline optimum over $p \in \Delta$ for feedback x. Therefore, if we can produce primal points x_t such that the feedback in each step is good wrt the dual p_t iterate, the framework of online optimization guarantees the average iterate has good offline optimum, i.e. $f(\overline{x})$. The formal result is given below:

Theorem 18. *Let $\{p_t\}$ be the iterates produced by MW or Exp Hedge when given feedback $\ell_t := b - Ax_t\rangle$ where x_t is a solution to the following*

$$x_t \in K, \quad \langle p_t, Ax_t - b \rangle \leq 0, \qquad \max_i \langle a_i, x_t \rangle - b_i \leq G.$$

Then in $T \lesssim G^2 \log n / \varepsilon^2$ iteration, we can produce an approximately feasible solution

$$\overline{x} := \frac{1}{T} \sum_{t=1}^{T} x_t \in K, \qquad A\overline{x} \leq b + \varepsilon 1_n.$$

Proof: The online algorithm gives us the following guarantee

$$f(\overline{x}) = \max_{p \in \Delta} \langle p, b - A\overline{x} \rangle \geq \frac{1}{T} \Big(\sum_{t \in [T]} \langle p_t, b - Ax_t \rangle - R_T \Big),$$

where R_T is the regret guarantee. Due to our oracle, we always find a solution x_t such that $\langle p_t, b - Ax_t \rangle$, and further the feedback is always bounded $b - Ax_t \leq G1_m$. Therefore our regret guarantee can be bounded

$$R_T \leq \sum_{t \in [T]} \frac{\eta G^2}{2} + \log n,$$

where we used $D(p \mid p_0) \leq \log n$ for $p_0 = 1_n/n$. Putting this together gives

$$f(\overline{x}) \geq 0 - \frac{\eta G^2}{2} - \frac{\log n}{\eta T} \gtrsim -\sqrt{\frac{G^2 \log n}{T}} \gtrsim -\varepsilon$$

where we chose $\eta \approx \sqrt{\log n / G^2 T}$ to balance terms and substituted $T \approx G^2 \log n / \varepsilon^2$. The result follows by definition of $f(\overline{x}) = \max_i (b - A\overline{x})_i$.

5.2 Application: Approximate Caratheodory Theorem

The following is a fundamental result in linear algebra.

Theorem 19 (Caratheodory's Theorem)**.** *Let $u \in \text{conv}\{v_1, ..., v_m\} \subseteq \mathbb{R}^d$, then u can be written as a convex combination of $\leq d+1$ vertices:*

$$\exists \lambda \in \Delta_m \quad \textit{with} \quad \textit{supp}(\lambda) \leq d+1 : \qquad u = V\lambda = \sum_{j \in [m]} \lambda_j v_j.$$

It turns out that if we allow some error, we can in fact give a much sparser representation.

Theorem 20 (Caratheodory's Theorem)**.** *Let $u \in K := \text{conv}\{v_1, ..., v_m\} \subseteq \mathbb{R}^d$ with $\max_j \|v_j\|_p \leq R$, then u can be approximately written as a convex combination of $T = O(p/\varepsilon^2)$ vertices:*

$$\exists \lambda \in \Delta_m \quad \textit{with} \quad \textit{supp}(\lambda) \leq T : \qquad \|u - V\lambda\|_p \leq \varepsilon R.$$

Proof: We will give an algorithmic proof of this result using Mirror descent. For the sake of analysis, we can shift by u and normalize so we are trying to represent $0 \in \text{conv}\{v_1, ..., v_m\}$ with $\max_j \|v_j\|_p \leq 1$. Note

$$\min_{\lambda \in \Delta_m} \|V\lambda\|_p = \min_{\lambda \in \Delta_m} \max_{\|y\|_q \leq 1} \langle y, V\lambda \rangle = \max_{\|y\|_q \leq 1} \min_{\lambda \in \Delta_m} \langle y, V\lambda \rangle =: \max_{\|y\|_q \leq 1} g(y),$$

where the first step was by duality of norms for $\frac{1}{p} + \frac{1}{q} = 1$ and the second was by Sion's Minimax Theorem. Note that g is a concave function of y since it is a minimum of affine functions.

Once again our plan is to use Mirror descent to keep track of the dual variable y. We can compute g if we have access to an optimization oracle for $K = \text{conv}\{v_1, ..., v_m\}$. Further, by the Envelope theorem it can be shown that $\partial g(y) \ni v_j$ for $v_j \in \arg\min_{v \in K} \langle y, v \rangle$. Specifically, we will use the regularizer $\phi(x) := \frac{1}{2(q-1)} \|x\|_q^2$, which can be shown to be 1-strongly convex wrt $\|\cdot\|_q$, along with constraint $y \in B_q$ and initial iterate $y_0 = 0$. Further, note by our boundedness assumption that this gives the Lipschitz-ness condition

$$\|\partial g(y)\|_p = \|v_j\|_p \leq 1,$$

where $\|\cdot\|_p$ is the dual norm used for the subgradient.

In order to find a good representation, it is enough in each iteration to find an update that is good for the proxy. Specifically, in each iteration we will use our optimization oracle to compute $v_{j_t} \in \arg\min_{v \in K} \langle y, v \rangle$ and let $\lambda_t := e_{j_t}$. Let $\overline{\lambda} := \frac{1}{T} \sum_{t \in [T]} \lambda_t$, then

$$\|V\overline{\lambda}\|_p = \max_{\|y\|_q \leq 1} \langle y, V\overline{\lambda} \rangle \leq \frac{1}{T} \Bigg(\sum_{t \in [T]} \langle y_t, V\lambda_t \rangle + R_T \Bigg),$$

where again R_T is the regret bound given by Mirror descent. Applying the Mirror Descent analysis from above, we get regret bound

$$R_T \leq \sum_{t \in [T]} \frac{\eta \|v_{j_t}\|_p^2}{2} + D_\phi(y \mid y_0) \leq \frac{\eta T}{2} + \frac{1}{2(q-1)},$$

where in the last step we used $\max_{y \in B_q} D_\phi(y \mid 0) \leq 1/2(q-1)$, as can be verified directly. Plugging this back in gives approximation guarantee

$$\|V\overline{\lambda}\|_p =\lesssim 0 + \eta + \frac{1}{\eta T(q-1)} \lesssim \sqrt{\frac{p}{T}} \lesssim \varepsilon,$$

where in the first step we used that $0 \in K$ so $\min_{v \in K} \langle y_t, v \rangle \leq 0$ for all iterations, and in the remainder we chose $\eta = 1/\sqrt{(q-1)T}$ to balance terms, used $\frac{1}{p} + \frac{1}{q} = 1$ and substituted $T \approx p/\varepsilon^2$.

Finally, note that in each iteration the oracle produces a single vertex of the polytope K as feedback. Therefore in T iterations, our output is the average of $\leq T \lesssim p/\varepsilon^2$ vertices, which gives our sparsity guarantee.