Online Convex Optimization and Mirror Descent

Yudong Chen

Reading:

- Chapter 21 of Duchi's notes
- Xinhua Zhang, short notes on mirror descent
- Elad Hazan, "Introduction to Online Convex Optimization"
- Section 4 of Bubeck's monograph
- Lectures 5–9 in Jiantao Jiao's course on convex optimization

1 Online Convex Optimization

The setup can be described as a two-player sequential game:

- Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a *convex* feasible set (we call it the *parameter space* in this lecture).
- At each time t, player 1 (the *learner*) chooses some $x_t \in \mathcal{X}$.
- Player 2 (the *adversary*, or *nature*) then chooses a *convex* loss function $f_t : \mathcal{X} \to \mathbb{R}$.

Note that the learner commits to x_t **before** seeing f_t, whereas the adversary may adapt its choice of f_t to $x_1, \ldots, x_t$. The goal for the learner is to minimize the average *regret*, defined as

$$\frac{1}{T}\sum_{t=1}^{T}\left(f_t(x_t) - f_t(x^*)\right),$$

where $x^* := \operatorname{argmin}_{x\in\mathcal{X}} \sum_{t=1}^{T} f_t(x)$ is the best *fixed* decision in hindsight. In general, we want the average regret to go to zero as $T \to \infty$.

1.1 Examples

Here are some examples of problems that fall into the framework of online convex optimization.

1. **Online support vector machine**: At each time t, the learner picks a vector $x_t \in \mathbb{R}^d$. Then, a data point $(a_t, y_t) \in \mathbb{R}^d \times \{\pm 1\}$ is revealed, and the learner incurs loss $f_t(x_t)$, where $f_t(x) = \max\{1 - y_t \langle x, a_t \rangle, 0\}$. (This loss function is called the *hinge loss*.)
2. **Online logistic regression**: Same setup, except now the loss function is $f_t(x) = \log\left(1 + e^{-y_t\langle x, a_t\rangle}\right)$. (This is the *logistic loss*.)

3. **Expert prediction/adversarial bandit**: There are d experts/arms. At each time t, each expert makes a prediction (for example "I predict the stock market will go up tomorrow"). At each time t, the learner chooses a weight vector $x_t = (x_{t1}, \ldots, x_{td})$, where

$$x_{tj} = \text{weight for expert } j = \text{probability of pulling arm } j.$$

The parameter space is $\mathcal{X} = \Delta_d := \{x \in \mathbb{R}^d : \sum_j x_j = 1, \, x_j \geq 0\}$, which is the probability simplex in $\mathbb{R}^d$. Then losses

$$l_{tj} = \mathbb{I}\{\text{expert } j \text{ is wrong at time } t\} = \text{loss of arm } j \text{ at time } t$$

are revealed for $j = 1, \ldots, d$, and the learner incurs expected/average loss $f_t(x_t) = \langle x_t, l_t \rangle$. Note that $\nabla f_t(x_t) = l_t$.

2 Online Gradient Descent

Gradient descent extends naturally to an algorithm for online convex optimization. Online gradient descent computes, at each iteration $t + 1$:

$$\begin{aligned} x_{t+1} &= P_{\mathcal{X}}(x_t - \alpha_t g_t) \\ &= \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \langle g_t, x \rangle + \frac{1}{2\alpha_t} \|x - x_t\|_2^2 \right\}. \end{aligned}$$

where α_t is the step size and $g_t = \nabla f_t(x_t)$. (This is can be generalized to the setting where f_t is non-differentiable, in which case $g_t \in \partial f_t(x_t)$ is a subgradient of f_t at x_t.)

3 Bregman Divergence

We will next see how to extend gradient descent to a more general algorithm. First, we need to introduce the notion of Bregman divergence. Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be a differentiable convex function.

Definition 1 (Bregman Divergence)**.** The **Bregman divergence** associated with ψ is a function $B_\psi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined by

$$B_\psi(x, y) := \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle$$

Remark 1. By the convexity of ψ, the Bregman divergence B_ψ is always non-negative. One may loosely think of $B_\psi(x, y)$ as a measure of "distance" between x and y; however, the Bregman divergence is not necessarily symmetric or need not satisfy the triangle inequality.

3.1 Examples

1. **Euclidean distance.** Let $\psi(x) = \frac{1}{2}\|x\|_2^2$. Then $B_\psi(x, y) = \frac{1}{2}\|x - y\|_2^2$.

2. **Mahalanobis distance.** Let $\psi(x) = \frac{1}{2}x^\top A x =: \frac{1}{2}\|x\|_A^2$, where $A \succcurlyeq 0$.
   Then $B_\psi(x, y) = \frac{1}{2}(x - y)^\top A (x - y) = \frac{1}{2}\|x - y\|_A^2$.

3. **KL-divergence.** Let $\psi(x) = \sum_{j=1}^d x_j \log x_j$ be the negative entropy. Note that ψ is convex on $\mathbb{R}_+^d$.
   Then $B_\psi(x, y) = \sum_{j=1}^d x_j \log \frac{x_j}{y_j} = D_{\text{KL}}(x, y)$ for all $x, y \in \Delta_d$, where $D_{\text{KL}}(\cdot, \cdot)$ is the Kullback-Leibler divergence, and Δ_d denotes the probability simplex in d-dimension, .

4 Online Mirror Descent (OMD)

This is a generalization of gradient descent using Bregman divergences. At iteration t:

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \langle g_t, x \rangle + \frac{1}{\alpha_t} B_\psi(x, x_t) \right\} \tag{1}$$

Remark 2. $\langle g_t, x \rangle + \frac{1}{\alpha_t} B_\psi(x, x_t)$ is convex in x. Hence this is a convex optimization problem.

4.1 Special cases of OMD

Gradient descent $\psi(x) = \frac{1}{2} \|x\|_2^2$

Exponentiated gradient descent This is online mirror descent with $\mathcal{X} = \Delta_d$, $\psi(x) = \sum_j x_j \log x_j$, and $B_\psi(x, y) = D_{\text{KL}}(x, y)$. At iteration t:

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \langle g_t, x \rangle + \frac{1}{\alpha_t} D_{\text{KL}}(x, x_t) \right\}.$$

To explicit calculate x_{t+1}, we write the Lagrangian:

$$L(x, \lambda, \tau) = \langle g_t, x \rangle + \frac{1}{\alpha} \sum_{i=1}^d x_i \log \frac{x_i}{x_{t,i}} - \langle \lambda, x \rangle + \tau \left(\langle \mathbb{I}, x \rangle - 1 \right).$$

Here, $\lambda \in \mathbb{R}^d$ is the multiplier for the element-wise constraint $x \geq 0$, and $\tau \in \mathbb{R}$ is the multiplier for the constraint $\langle \mathbb{I}, x \rangle = 1$. Taking $\frac{\partial}{\partial x} L(x, \lambda, \tau) = 0$ gives

$$x_{t+1,i} = x_{t,i} \exp\left(-\alpha g_{t,i} + \lambda_i \alpha - \tau \alpha - 1 \right) > 0.$$

Hence the constraint $x \geq 0$ is inactive, which implies $\lambda = \vec{0}$. We choose τ to normalize x, giving

$$x_{t+1,i} = \frac{x_{t,i} \exp(-\alpha g_{t,i})}{Z_t} \qquad \text{where } Z_t = \sum_{j=1}^d x_{t,j} \exp\left(-\alpha g_{t,j}\right) \tag{2}$$

$$= \frac{\exp\left(-\sum_{k=1}^t \alpha_k g_{k,i}\right)}{\text{normalization-factor}}. \tag{3}$$

We sometimes write this as

$$x_{t+1} = \text{soft-argmin} \left\{ \sum_{k=1}^t \alpha_k g_{k,i}, \ i = 1, \ldots, d \right\}. \tag{4}$$

Remark 3. In the context of the expert problem, $g_{k,i}$ is the loss of expert i at time k. Hence, $\sum_{k=1}^t g_{k,i}$ is the total loss of expert i up to time t. Hence exponentiated gradient descent favors experts with low historical loss, but still assigns positive weight to every expert. This algorithm can thus be interpreted as a smoothed version of "follow the leader", where the weights are updated in an multiplicative fashion. (Variants of) exponentiated gradient descent is also known as **multiplicative weight update** (MWU), **follow-the-regularized-leader** (FTRL), **fictitious play** (FP), **Hedge algorithm**, and **entropic mirror descent**.

5 Analysis of Online Mirror Descent

We recall some definitions.

Definition 2 (Strong convexity)**.** ψ is 1-*strongly convex* with respect to $\|\cdot\|$ if , for all y, x:

$$\psi(x) - \psi(y) - \langle g, x-y \rangle \geq \frac{1}{2}\|x-y\|^2, \quad \text{for all } g \in \partial\psi(y).$$

This is equivalent to $B_\psi(x,y) \geq \frac{1}{2}\|x-y\|^2$ by definition of Bregman divergence.

Example 1. Let $\psi(x) = \sum_j x_j \log x_j$ be negative entropy. Then by *Pinsker's inequality*, we have

$$B_\psi(x,y) = D_{\mathrm{KL}}(x,y) \geq \frac{1}{2}\|x-y\|_1^2. \tag{5}$$

In other words, the negative entropy is 1-strongly convex with respect to the ℓ_1 norm.

Definition 3 (Dual norm)**.** The dual norm of $\|\cdot\|$ is the norm $\|\cdot\|_*$ defined by

$$\|y\|_* = \sup_{x:\|x\|\leq 1} \langle x, y \rangle.$$

Example 2. The dual norm of $\|\cdot\|_2$ is $\|\cdot\|_2$. The dual norm of $\|\cdot\|_\infty$ is $\|\cdot\|_1$. The dual norm of $\|\cdot\|_{\mathrm{nuc}}$ (nuclear norm) is $\|\cdot\|_{\mathrm{op}}$ (operator norm).

Theorem 1. *Suppose that ψ is 1-strongly convex with respect to $\|\cdot\|$ with dual norm $\|\cdot\|_*$. Then online mirror descent (1) with constant step size $\alpha_t \equiv \alpha$ satisfies the regret bound*

$$\frac{1}{T}\sum_{t=1}^{T}\left[f_t(x_t) - f_t(x^*)\right] \leq \frac{1}{\alpha T}B_\psi(x^*, x_1) + \frac{\alpha}{2T}\sum_{t=1}^{T}\|g_t\|_*^2.$$

Proof. Recall that $x_{t+1} = \operatorname{argmin}_{x\in\mathcal{X}}\left\{\langle g_t, x\rangle + \frac{1}{\alpha}B_\psi(x, x_t)\right\}$. By the optimality condition for constrained optimization (negative gradient lies in the normal cone), we have

$$\begin{aligned}
0 &\leq \left\langle g_t + \frac{1}{\alpha}\frac{\partial}{\partial x}B_\psi(x,x_t)\Big|_{x=x_{t+1}}, x^* - x_{t+1}\right\rangle \\
&= \left\langle g_t + \frac{1}{\alpha}\left(\nabla\psi(x_{t+1}) - \nabla\psi(x_t)\right), x^* - x_{t+1}\right\rangle.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
f_t(x_t) - f_t(x^*) &\leq \langle g_t, x_t - x^*\rangle && \text{convexity of } f_t \\
&= \langle g_t, x_{t+1} - x^*\rangle + \langle g_t, x_t - x_{t+1}\rangle \\
&\leq \frac{1}{\alpha}\langle \nabla\psi(x_{t+1}) - \nabla\psi(x_t), x^* - x_{t+1}\rangle + \langle g_t, x_t - x_{t+1}\rangle && \text{last display equation} \\
&= \frac{1}{\alpha}\left[B_\psi(x^*, x_t) - B_\psi(x^*, x_{t+1}) - B_\psi(x_{t+1}, x_t)\right] + \langle g_t, x_t - x_{t+1}\rangle,
\end{aligned}$$

where the last step follows from direct calculation using definition and is sometimes known as the "three-point identity" for Bregman divergence (HW2 Q3.3). Let us sum over $t = 1, \ldots, T$. The sum telescopes and simplifies to

$$\begin{aligned}\sum_{t=1}^{T} (f_t(x_t) - f_t(x^*)) &\le \frac{1}{\alpha} \left[B_\psi(x^*, x_1) - B_\psi(x^*, x_{T+1})\right] + \sum_{t=1}^{T} \left[-\frac{1}{\alpha} B_\psi(x_{t+1}, x_t) + \langle g_t, x_t - x_{t+1}\rangle\right] \\ &\le \frac{1}{\alpha} B_\psi(x^*, x_1) + \sum_{t=1}^{T} \left[-\frac{1}{\alpha} B_\psi(x_{t+1}, x_t) + \langle g_t, x_t - x_{t+1}\rangle\right]\end{aligned}$$

To control the last RHS term, we observe that

$$\begin{aligned}\langle g_t, x_t - x_{t+1}\rangle &\le \|g_t\|_* \|x_t - x_{t+1}\| && \text{definition of dual norm} \\ &\le \frac{\alpha}{2} \|g_t\|^2 + \frac{1}{2\alpha} \|x_t - x_{t+1}\|^2 && ab \le \frac{1}{2}(a^2 + b^2) \\ &\le \frac{\alpha}{2} \|g_t\|_*^2 + \frac{1}{\alpha} B_\psi(x_{t+1}, x_t) && \text{strong convexity of } \psi.\end{aligned}$$

Combining pieces, we obtain

$$\sum_{t=1}^{T} (f_t(x_t) - f_t(x^*)) \le \frac{1}{\alpha} B_\psi(x^*, x_1) + \frac{\alpha}{2} \sum_{t=1}^{T} \|g_t\|_*^2.$$

Dividing both sides by $\frac{1}{T}$ gives the desired regret bound. □

6 Applications

6.1 Online (sub)-gradient descent

Let $\psi(x) = \frac{1}{2} \|x\|_2^2$. Then ψ is 1-strongly convex with respect to $\|\cdot\|_2$, and the dual norm is $\|\cdot\|_2$. Suppose each f_t is L-Lipschitz, which implies $\|g_t\|_2 \le M$. Then the regret bound is

$$\frac{1}{T} \sum_{t=1}^{T} (f_t(x_t) - f_t(x^*)) \le \frac{1}{2\alpha T} \|x^* - x_1\|_2^2 + \frac{\alpha}{2T} T \cdot M^2.$$

Choosing $\alpha = \frac{\|x^* - x_1\|_2}{M\sqrt{T}}$ to minimize the RHS gives

$$\frac{1}{T} \sum_{t=1}^{T} (f_t(x_t) - f_t(x^*)) \le \frac{\|x^* - x_1\|_2 M}{\sqrt{T}}.$$

Remark 4. The above bound implies an $O(\frac{1}{\sqrt{T}})$ convergence rate for the offline setting where all $f_t \equiv f$. In particular, letting $\bar{x} = \frac{1}{T} \sum_{t=1}^{T} x_t$, we have

$$f(\bar{x}) - f(x^*) \le \frac{1}{T} \sum_{t=1}^{T} [f(x_t) - f(x^*)] \le \frac{\|x^* - x_1\|_2 M}{\sqrt{T}},$$

where the first step above is by Jensen's inequality. This recovers the result from Lecture 17 on subgradient descent.

6.2 Exponentiated gradient descent

Let $\mathcal{X} = \Delta_d$, and $\psi(x) = \sum_j x_j \log x_j$ be the negative entropy. Then ψ is 1-strongly convex with respect to $\|\cdot\|_1$, with dual norm $\|\cdot\|_\infty$. Then the regret bound is

$$\frac{1}{T}\sum_{t=1}^{T}\left(f_t(x_t) - f_t(x^*)\right) \le \frac{1}{\alpha T} D_{\mathrm{KL}}\left(x^*, x_1\right) + \frac{\alpha}{2T}\sum_{t=1}^{T}\|g_t\|_\infty^2 .$$

If in addition we take the initial iterate $x_1 = (\frac{1}{d}, \ldots, \frac{1}{d})$ to be the uniform distribution, then one can verify that $D_{\mathrm{KL}}\left(x^*, x_1\right) \le \log d$. Also, set $\alpha = \sqrt{\frac{\log d}{2T \max_t \|g_t\|_\infty^2}}$. Then the average regret is

$$\frac{1}{T}\sum_{t=1}^{T}\left(f_t(x_t) - f_t(x^*)\right) \le \sqrt{\frac{\log d \cdot \max_t \|g_t\|_\infty^2}{T}}. \tag{6}$$

Remark 5. Compared to online gradient descent, the dependence on the gradients g_t is $\max_t \|g_t\|_\infty$ instead of $\max_t \|g_t\|_2$. Thus exponentiated gradient descent can do better than gradient descent when the gradients g_t are small in magnitude and not sparse.

6.3 Expert problem

Recall that l_{tj} is the loss of expert j at time t, and that $g_t = l_t \in \{0,1\}^d$. Thus $\|g_t\|_\infty \le 1$. Plugging this into the bound for exponentiated gradient descent gives

$$\frac{1}{T}\sum_{t=1}^{T}\left(f_t(x_t) - f_t(x^*)\right) \le \sqrt{\frac{\log d}{T}}$$

Remark 6. This regret bound is optimal for the expert problem. In comparison, gradient descent would get $\sqrt{\frac{d}{T}}$ regret, which has an exponentially larger dependence on the dimension d.

7 Extensions

1. We chose our step size α to be proportional to $\frac{1}{\sqrt{T}}$. This requires the time horizon to be known to the algorithm. If T is not known, one can use a varying step size $\alpha_t = \frac{1}{\sqrt{t}}$ and prove essentially the same guarantees (under a slightly stronger boundedness assumption; see Duchi's notes.)

2. **Improved bounds.** If more is known about the loss function f_t, then better regret bounds (in the online setting) and convergence rates (in the offline setting) can be obtained.

   - f_t is smooth (gradient is Lipschitz): We have an improvement $\frac{1}{\sqrt{T}} \to \frac{1}{T}$ in average regret. This can be further improved to $\frac{1}{T^2}$ using ideas similar to Nesterov's acceleration.
   - f_t is strongly convex: We have an improvement $\frac{1}{\sqrt{T}} \to \frac{\log T}{T}$ in average regret.

   See Xinhua Zhang's notes for details.

3. So far, we assumed that we observe the losses of *all* the experts/arms, even those we did not choose/pull. This is the *full information* setting. Next week, we will look at the "bandit information" setting, where we only observe the loss of the expert/arm that we choose/pull, that is, we only see one entry of $\nabla f_t = g_t = l_t$.

8 Why is it called mirror descent?

The online mirror descent update (1) can be written equivalently as

$$\text{compute} \quad y_{t+1} \in \mathbb{R}^d \quad \text{such that } \nabla\psi(y_{t+1}) = \nabla\psi(x_t) - \alpha_t g_t \tag{7a}$$
$$\text{compute} \quad x_{t+1} \in \operatorname*{argmin}_{x\in\mathcal{X}} B_\psi(x, y_{t+1}). \tag{7b}$$

To see the equivalence, we continue from (7b) to get

$$\begin{aligned} x_{t+1} &= \operatorname*{argmin}_{x\in\mathcal{X}} \{\psi(x) - \psi(y_{t+1}) - \langle \nabla\psi(y_{t+1}), x - y_{t+1}\rangle\} \\ &= \operatorname*{argmin}_{x\in\mathcal{X}} \{\psi(x) - \psi(y_{t+1}) - \langle \nabla\psi(x_t) - \alpha_t g_t, x - y_{t+1}\rangle\} && \text{by (7a)} \\ &= \operatorname*{argmin}_{x\in\mathcal{X}} \{\alpha_t \langle g_t, x\rangle + \psi(x) - \langle \nabla\psi(x_t), x\rangle\} && \text{omit terms independent of } x \\ &= \operatorname*{argmin}_{x\in\mathcal{X}} \{\alpha_t \langle g_t, x\rangle + B_\psi(x, x_t)\} = (1). \end{aligned}$$

One can view $\nabla\psi : \mathbb{R}^d \to \mathbb{R}^d$ as a mapping from the primal space to the dual/mirror space, and $(\nabla\psi)^{-1}$ is the inverse mapping from the mirror space to the primal space. Therefore, in (7a), we first map x_t to $\nabla\psi(x_t)$, then perform an gradient descent step $\nabla\psi(x_t) - \alpha_t g_t$ in this mirror space, and finally map back to the primal space to obtain $y_{t+1} = (\nabla\psi)^{-1}(\nabla\psi(x_t) - \alpha_t g_t)$. The update in (7b) can be viewed as the projection of y_{t+1} to $\mathcal{X}$ with respect to the Bregman divergence B_ψ.

For an illustration see the following plot from Bubeck.

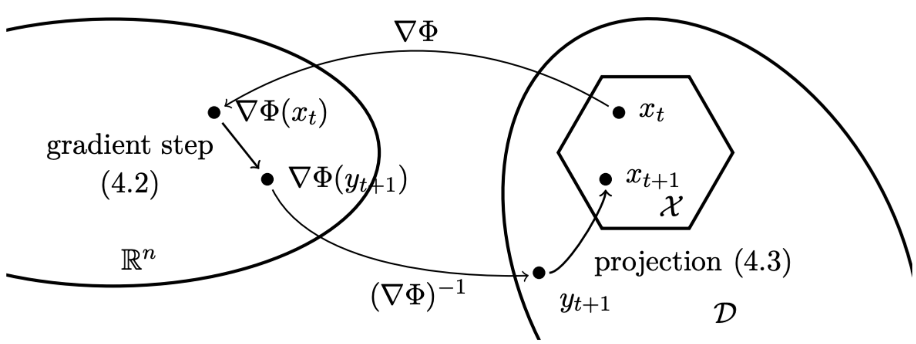

Figure 4.1: Illustration of mirror descent.

9 Lazy mirror descent

The above perspective suggests a somewhat more efficient variant of mirror descent, where we use y_t instead of x_t on the RHS of (7a). This is called *lazy mirror descent*, as given below:

$$\text{compute} \quad y_{t+1} \in \mathbb{R}^d \quad \text{such that } \nabla\psi(y_{t+1}) = \nabla\psi(y_t) - \alpha_t g_t \tag{8a}$$
$$\text{compute} \quad x_{t+1} \in \operatorname*{argmin}_{x\in\mathcal{X}} B_\psi(x, y_{t+1}). \tag{8b}$$

Note that the step (8a) is equivalent to

$$\theta_{t+1} = \nabla\psi(y_1) - \sum_{t=1}^{\infty} \alpha_t g_t, \tag{9}$$

$$y_{t+1} \in (\nabla\psi)^{-1}(\theta_{t+1}). \tag{10}$$

Here we are averaging the g_t's in the dual space. Therefore, lazy mirror descent is also known as (Nesterov's) *Dual Averaging*.

In the original mirror descent (7), we go back and forth between the primal and mirror space: $x_t \to \nabla\psi(x_t) \to (\nabla\psi)^{-1}(\nabla\psi(x_t) - \alpha_t g_t)$. In lazy mirror descent, the step (8b) or (9) is done purely in the mirror space; only when asked to output x_{t+1}, we map the dual point θ_{t+1} back to the primal space. One may notice that if $g_t = \nabla f_t(x_t)$ is the gradient at x_t, then one needs to compute the primal points y_t and x_t in every iteration. However, this only involves the backward map $\nabla^{-1}\psi$, so we do not need to compute the forward map $\nabla\psi$ as in the original mirror descent. This can be advantageous in the distributed setting, or when $\nabla^{-1}\psi$ is easier to compute than $\nabla\psi$.

In the special case of $\mathcal{X} = \Delta_d$ and $\psi(x) = \sum_j x_j \log x_j$ (i.e., exponentiated gradient descent), mirror descent and lazy mirror descent are equivalent, corresponding to the updates (2) and (3) respectively.

9.1 Regret bound

Lazy mirror descent enjoys a similar convergence guarantee as mirror descent. Recall that each f_t is convex and $g_t = \nabla f_t(x_t)$.

Theorem 2. *Suppose that ψ is 1-strongly convex with respect to $\|\cdot\|$ with dual norm $\|\cdot\|_*$. Consider the lazy mirror descent (8) with constant step size $\alpha_t \equiv \alpha$ and initial point $x_1 = y_1$ satisfying $\nabla\psi(y_1) = 0$. We have the regret bound*

$$\frac{1}{T}\sum_{t=1}^{T}\left[f_t(x_t) - f_t(x^*)\right] \le \frac{1}{\alpha T}\left(\psi(x^*) - \psi(x_1)\right) + \frac{2\alpha}{T}\sum_{t=1}^{T}\|g_t\|_*^2.$$

Proof. For each t define the potential function $L_t(x) := \alpha\sum_{s=1}^{t-1}\langle g_s, x\rangle + \psi(x)$, which is 1-strongly convex since ψ is. From (9) and $\nabla\psi(y_1) = 0$, we have

$$x_t \in \operatorname*{argmin}_{x\in\mathcal{X}} B_\psi(x, y_t) = \operatorname*{argmin}_{x\in\mathcal{X}} \psi(x) - \langle\nabla\psi(y_t), x\rangle = \operatorname*{argmin}_{x\in\mathcal{X}} L_t(x).$$

By strong convexity we have

$$L_{t+1}(x_{t+1}) - L_{t+1}(x_t) \le \langle\nabla L_{t+1}(x_{t+1}), x_{t+1} - x_t\rangle - \frac{1}{2}\|x_{t+1} - x_t\|^2 \le -\frac{1}{2}\|x_{t+1} - x_t\|^2,$$

where the last step follows from the first-order optimality condition for x_{t+1} w.r.t. L_{t+1}. We also have

$$L_{t+1}(x_{t+1}) - L_{t+1}(x_t) = L_t(x_{t+1}) - L_t(x_t) + \alpha\langle g_t, x_{t+1} - x_t\rangle \ge \alpha\langle g_t, x_{t+1} - x_t\rangle$$

by optimality of x_t w.r.t. L_t. Combining the last two inequalities, we get

$$\frac{1}{2}\|x_{t+1} - x_t\|^2 \le -\alpha\langle g_t, x_{t+1} - x_t\rangle \le \alpha\|g_t\|_*\|x_{t+1} - x_t\|.$$

This implies that $\|x_{t+1} - x_t\| \leq 2\alpha \|g_t\|_*$ and thus

$$\langle g_t, x_t - x_{t+1} \rangle \leq \|g_t\|_* \|x_{t+1} - x_t\| \leq 2\alpha \|g_t\|_*^2. \tag{11}$$

We claim that

$$\sum_{t=1}^{T-1} \langle g_t, x_{t+1} \rangle + \frac{\psi(x_1)}{\alpha} \leq \sum_{t=1}^{T-1} \langle g_t, x \rangle + \frac{\psi(x)}{\alpha}, \qquad \forall x \in \mathcal{X}. \tag{12}$$

We prove by induction on T. For $T = 1$, the inequality (12) becomes $\psi(x_1) \leq \psi(x^*)$, which holds because x_1 satisfies $\nabla\psi(x_1) = 0$ and is thus a minimizer of ψ. Now assume that the bound (12) holds for some T. Setting $x = x_{T+1}$ we get

$$\sum_{t=1}^{T-1} \langle g_t, x_{t+1} \rangle + \frac{\psi(x_1)}{\alpha} \leq \sum_{t=1}^{T-1} \langle g_t, x_{T+1} \rangle + \frac{\psi(x_{T+1})}{\alpha}.$$

Hence

$$\sum_{t=1}^{T} \langle g_t, x_{t+1} \rangle + \frac{\psi(x_1)}{\alpha} \leq \underbrace{\langle g_T, x_{T+1} \rangle + \sum_{t=1}^{T-1} \langle g_t, x_{T+1} \rangle + \frac{\psi(x_{T+1})}{\alpha}}_{L_{T+1}(x_{T+1})} \leq \underbrace{\sum_{t=1}^{T} \langle g_t, x \rangle + \frac{\psi(x)}{\alpha}}_{L_{T+1}(x)},$$

where the last step holds since $x_{T+1} \in \operatorname{argmin}_{x \in \mathcal{X}} L_{T+1}(x)$. This proves (12) for $T + 1$.

Combining pieces, we obtain

$$\begin{aligned}
\sum_{t=1}^{T} [f_t(x_t) - f_t(x^*)] &\leq \sum_{t=1}^{T-1} \langle g_t, x_t - x^* \rangle && f_t \text{ is convex} \\
&= \sum_{t=1}^{T-1} \langle g_t, x_t - x_{t+1} \rangle + \sum_{t=1}^{T-1} \langle g_t, x_{t+1} - x^* \rangle && \text{(11), and (12) with x=}x^*, \\
&\leq \sum_{t=1}^{T-1} 2\alpha \|g_t\|_*^2 + \frac{\psi(x^*) - \psi(x_1)}{\alpha}.
\end{aligned}$$

Dividing both sides by T proves Theorem 2. □