Bregman Divergence and Mirror Descent

1 Bregman Divergence

Motivation

- Generalize squared Euclidean distance to a class of distances that all share similar properties
- Lots of applications in machine learning, clustering, exponential family

Definition 1 (Bregman divergence) *Let $\psi : \Omega \to \mathbb{R}$ be a function that is: a) strictly convex, b) continuously differentiable, c) defined on a closed convex set Ω. Then the Bregman divergence is defined as*

$$\Delta_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y \rangle, \qquad \forall\, x, y \in \Omega. \tag{1}$$

That is, the difference between the value of ψ at x and the first order Taylor expansion of ψ around y evaluated at point x.

Examples

- Euclidean distance. Let $\psi(x) = \frac{1}{2}\|x\|^2$. Then $\Delta_\psi(x, y) = \frac{1}{2}\|x - y\|^2$.
- $\psi(x) = \sum_i x_i \log x_i$ and $\Omega = \left\{x \in \mathbb{R}^n_+ : \mathbf{1}'x = 1\right\}$, where $\mathbf{1} = (1, 1, \ldots, 1)'$. Then $\Delta_\psi(x, y) = \sum_i x_i \log \frac{x_i}{y_i}$ for $x, y \in \Omega$. This is called relative entropy, or Kullback–Leibler divergence between probability distributions x and y.
- L_p norm. Let $p \geq 1$ and $\frac{1}{p} + \frac{1}{q} = 1$. $\psi(x) = \frac{1}{2}\|x\|_q^2$. Then $\Delta_\psi(x, y) = \frac{1}{2}\|x\|_q^2 + \frac{1}{2}\|y\|_q^2 - \left\langle x, \nabla\frac{1}{2}\|y\|_q^2 \right\rangle$. Note $\frac{1}{2}\|y\|_q^2$ is not necessarily continuously differentiable, which makes this case not precisely consistent with our definition.

1.1 Properties of Bregman divergence

- Strict convexity in the first argument x. Trivial by the strict convexity of ψ.
- Nonnegativity: $\Delta_\psi(x, y) \geq 0$ for all x, y. $\Delta_\psi(x, y) = 0$ if and only if $x = y$. Trivial by strict convexity.
- Asymmetry: in general, $\Delta_\psi(x, y) \neq \Delta_\psi(y, x)$. Eg, KL-divergence. Symmetrization not always useful.
- Non-convexity in the second argument. Let $\Omega = [1, \infty)$, $\psi(x) = -\log x$. Then $\Delta_\psi(x, y) = -\log x + \log y + \frac{x-y}{y}$. One can check its second order derivative in y is $\frac{1}{y^2}(\frac{2x}{y} - 1)$, which is negative when $2x < y$.
- Linearity in ψ. For any $a > 0$, $\Delta_{\psi + a\varphi}(x, y) = \Delta_\psi(x, y) + a\Delta_\varphi(x, y)$.
- Gradient in x: $\frac{\partial}{\partial x}\Delta_\psi(x, y) = \nabla\psi(x) - \nabla\psi(y)$. Gradient in y is trickier, and not commonly used.
- Generalized triangle inequality:

$$\begin{aligned} \Delta_\psi(x, y) + \Delta_\psi(y, z) &= \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y \rangle + \psi(y) - \psi(z) - \langle \nabla\psi(z), y - z \rangle & (2)\\ &= \Delta_\psi(x, z) + \langle x - y, \nabla\psi(z) - \nabla\psi(y) \rangle. & (3) \end{aligned}$$

- Special case: ψ is called strongly convex with respect to some norm with modulus σ if

$$\psi(x) \geq \psi(y) + \langle \nabla\psi(y), x - y \rangle + \frac{\sigma}{2}\|x - y\|^2. \tag{4}$$

  Note the norm here is not necessarily the Euclidean norm. When the norm is Euclidean, this condition is equivalent to $\psi(x) - \frac{\sigma}{2}\|x\|^2$ being convex. For example, the $\psi(x) = \sum_i x_i \log x_i$ used in KL-divergence is 1-strongly convex over the simplex $\Omega = \left\{x \in \mathbb{R}^n_+ : \mathbf{1}'x = 1\right\}$, with respect to the L_1 norm (not so trivial). When ψ is σ strongly convex, we have

$$\Delta_\psi(x, y) \geq \frac{\sigma}{2} \|x - y\|^2 . \tag{5}$$

Proof: By definition $\Delta_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y\rangle \geq \frac{\sigma}{2} \|x - y\|^2$. ■

- Duality. Suppose ψ is strongly convex. Then

$$(\nabla\psi^*)(\nabla\psi(x)) = x, \qquad \Delta_\psi(x, y) = \Delta_{\psi^*} (\nabla\psi(y), \nabla\psi(x)) . \tag{6}$$

Proof: (for the first equality only) Recall

$$\psi^*(y) = \sup_{z\in\Omega} \{\langle z, y\rangle - \psi(z)\} . \tag{7}$$

sup must be attainable because ψ is strongly convex and Ω is closed. x is a maximizer if and only if $y = \nabla\psi(x)$. So

$$\psi^*(y) + \psi(x) = \langle x, y\rangle \quad \Leftrightarrow \quad y = \nabla\psi(x). \tag{8}$$

Since $\psi = \psi^{**}$, so $\psi^*(y) + \psi^{**}(x) = \langle x, y\rangle$, which means y is the maximizer in

$$\psi^{**}(x) = \sup_z \{\langle x, z\rangle - \psi^*(z)\} . \tag{9}$$

This means $x = \nabla\psi^*(y)$. To summarize, $(\nabla\psi^*)(\nabla\psi(x)) = x$. ■

- Mean of distribution. Suppose U is a random variable over an open set S with distribution μ. Then

$$\min_{x\in S} \ \mathbb{E}_{U\sim\mu}[\Delta_\psi(U, x)] \tag{10}$$

is optimized at $\bar{u} := \mathbb{E}_\mu[U] = \int_{u\in S} u\mu(u)$.

Proof: For any $x \in S$, we have

$$\begin{aligned}
&\mathbb{E}_{U\sim\mu}[\Delta_\psi(U, x)] - \mathbb{E}_{U\sim\mu}[\Delta_\psi(U, \bar{u})] && (11)\\
&= \mathbb{E}_\mu[\psi(U) - \psi(x) - (U - x)'\nabla\psi(x) - \psi(U) + \psi(\bar{u}) + (U - \bar{u})'\nabla\psi(\bar{u})] && (12)\\
&= \psi(\bar{u}) - \psi(x) + x'\nabla\psi(x) - \bar{u}'\nabla\psi(\bar{u}) + \mathbb{E}_\mu[-U'\nabla\psi(x) + U'\nabla\psi(\bar{u})] && (13)\\
&= \psi(\bar{u}) - \psi(x) - (\bar{u} - x)'\nabla\psi(x) && (14)\\
&= \Delta_\psi(\bar{u}, x). && (15)
\end{aligned}$$

This must be nonnegative, and is 0 if and only if $x = \bar{u}$. ■

- Pythagorean Theorem. If x^* is the projection of x_0 onto a convex set $C \subseteq \Omega$:

$$x^* = \operatorname*{argmin}_{x\in C} \Delta_\psi(x, x_0). \tag{16}$$

Then for all $y \in C$,

$$\Delta_\psi(y, x_0) \geq \Delta_\psi(y, x^*) + \Delta_\psi(x^*, x_0). \tag{17}$$

In Euclidean case, it means the angle $\angle yx^*x_0$ is obtuse. More generally

Lemma 2 *Suppose L is a proper convex function whose domain is an open set containing C. L is not necessarily differentiable. Let x^* be*

$$x^* = \operatorname*{argmin}_{x\in C} \ \{L(x) + \Delta_\psi(x^*, x_0)\} . \tag{18}$$

Then for any $y \in C$ we have

$$L(y) + \Delta_\psi(y, x_0) \geq L(x^*) + \Delta_\psi(x^*, x_0) + \Delta_\psi(y, x^*). \tag{19}$$

The projection in (16) is just a special case of $L = 0$. This property is the key to the analysis of many optimization algorithms using Bregman divergence.

Proof: Denote $J(x) = L(x) + \Delta_\psi(x, x_0)$. Since x^* minimizes J over C, there must exist a subgradient $d \in \partial J(x^*)$ such that

$$\langle d, x - x^* \rangle \geq 0, \qquad \forall\, x \in C. \tag{20}$$

Since $\partial J(x^*) = \{g + \nabla_{x=x^*}\Delta_\psi(x, x_0) : g \in \partial L(x^*)\} = \{g + \nabla\psi(x^*) - \nabla\psi(x_0) : g \in \partial L(x^*)\}$. So there must be a subgradient $g \in L(x^*)$ such that

$$\langle g + \nabla\psi(x^*) - \nabla\psi(x_0), x - x^* \rangle \geq 0, \qquad \forall\, x \in C. \tag{21}$$

Therefore using the property of subgradient, we have for all $y \in C$ that

$$\begin{aligned}
L(y) &\geq L(x^*) + \langle g, y - x^* \rangle && (22)\\
&\geq L(x^*) + \langle \nabla\psi(x_0) - \nabla\psi(x^*), y - x^* \rangle && (23)\\
&= L(x^*) - \langle \nabla\psi(x_0), x^* - x_0 \rangle + \psi(x^*) - \psi(x_0) && (24)\\
&\qquad + \langle \nabla\psi(x_0), y - x_0 \rangle - \psi(y) + \psi(x_0) && (25)\\
&\qquad - \langle \nabla\psi(x^*), y - x^* \rangle + \psi(y) - \psi(x^*) && (26)\\
&= L(x^*) + \Delta_\psi(x^*, x_0) - \Delta_\psi(y, x_0) + \Delta_\psi(y, x^*). && (27)
\end{aligned}$$

Rearranging completes the proof. ■

2 Mirror Descent for Batch Optimization

Suppose we want to minimize a function f over a set C. Recall the subgradient descent rule

$$x_{k+\frac{1}{2}} = x_k - \eta_k g_k, \quad \text{where} \quad g_k \in \partial f(x_k) \tag{28}$$

$$x_{k+1} = \operatorname*{argmin}_{x \in C} \frac{1}{2} \left\| x - x_{k+\frac{1}{2}} \right\|^2 = \operatorname*{argmin}_{x \in C} \frac{1}{2} \left\| x - (x_k - \eta_k g_k) \right\|^2. \tag{29}$$

This can be interpreted as follows. First approximate f around x_k by a first-order Taylor expansion

$$f(x) \approx f(x_k) + \langle g_k, x - x_k \rangle. \tag{30}$$

Then penalize the displacement by $\frac{1}{2\eta_k}\|x - x_k\|^2$. So the update rule is to find a regularized minimizer of the model

$$x_{k+1} = \operatorname*{argmin}_{x \in C} \left\{ f(x_k) + \langle g_k, x - x_k \rangle + \frac{1}{2\eta_k}\|x - x_k\|^2 \right\}. \tag{31}$$

It is trivial to see this is exactly equivalent to (29).

Mirror descent extension To generalize (31) beyond Euclidean distance, it is straightforward to use the Bregman divergence as a measure of displacement:

$$\begin{aligned}
x_{k+1} &= \operatorname*{argmin}_{x \in C} \left\{ f(x_k) + \langle g_k, x - x_k \rangle + \frac{1}{\eta_k}\Delta_\psi(x, x_k) \right\} && (32)\\
&= \operatorname*{argmin}_{x \in C} \left\{ \eta_k f(x_k) + \eta_k \langle g_k, x - x_k \rangle + \Delta_\psi(x, x_k) \right\}. && (33)
\end{aligned}$$

It is again equivalent to two steps:

$$\begin{aligned}
x_{k+\frac{1}{2}} &= \operatorname*{argmin}_{x} \left\{ f(x_k) + \langle g_k, x - x_k \rangle + \frac{1}{\eta_k}\Delta_\psi(x, x_k) \right\} && (34)\\
x_{k+1} &= \operatorname*{argmin}_{x \in C} \Delta_\psi(x, x_{k+\frac{1}{2}}). && (35)
\end{aligned}$$

The first order optimality condition for (34) is

$$\begin{aligned}
& g_k + \frac{1}{\eta_k}(\nabla\psi(x_{k+\frac{1}{2}}) - \nabla\psi(x_k)) = 0 && (36)\\
\iff \quad & \nabla\psi(x_{k+\frac{1}{2}}) = \nabla\psi(x_k) - \eta_k g_k && (37)\\
\iff \quad & x_{k+\frac{1}{2}} = (\nabla\psi)^{-1}(\nabla\psi(x_k) - \eta_k g_k) = (\nabla\psi^*)(\nabla\psi(x_k) - \eta_k g_k). && (38)
\end{aligned}$$

For example, in KL-divergence over simplex, the update rule becomes

$$x_{k+\frac{1}{2}}(i) = x_k(i)\exp(-\eta_k g_k(i)). \tag{39}$$

2.1 Rate of convergence for subgradient descent with Euclidean distance

We now analyze the rates of convergence of subgradient descent as in (31) and (33). It takes four steps.

1. Bounding on a single update

$$\|x_{k+1} - x^*\|_2^2 \le \left\|x_{k+\frac{1}{2}} - x^*\right\|^2 = \|x_k - \eta_k g_k - x^*\|_2^2 \quad (\le \text{ by the Pythagorean theorem in (17)}) \tag{40}$$

$$= \|x_k - x^*\|_2^2 - 2\eta_k \langle g_k, x_k - x^* \rangle + \eta_k^2 \|g_k\|_2^2 \tag{41}$$

$$\le \|x_k - x^*\|_2^2 - 2\eta_k (f(x_k) - f(x^*)) + \eta_k^2 \|g_k\|_2^2 . \tag{42}$$

2. Telescope over $k = 1, \ldots, T$ (summing them up):

$$\|x_{T+1} - x^*\|_2^2 \le \|x_1 - x^*\|_2^2 - 2\sum_{k=1}^{T} \eta_k (f(x_k) - f(x^*)) + \sum_{k=1}^{T} \eta_k^2 \|g_k\|_2^2 . \tag{43}$$

3. Bounding by $\|g_k\|_2^2 \le G^2$ and $\|x_1 - x^*\|_2^2 \le R^2 := \max_{x \in C} \|x_1 - x\|_2^2$:

$$2\sum_{k=1}^{T} \eta_k (f(x_k) - f(x^*)) \le R^2 + G^2 \sum_{k=1}^{T} \eta_k^2. \tag{44}$$

4. Denote $\epsilon_k = f(x_k) - f(x^*)$ and rearrange

$$\min_{k \in \{1,\ldots,T\}} \epsilon_k \le \frac{R^2 + G^2 \sum_{k=1}^{T} \eta_k^2}{2\sum_{k=1}^{T} \eta_k}. \tag{45}$$

Denote $[T] := \{1, 2, \ldots, T\}$. By setting the step size $\eta_k = \frac{R}{G\sqrt{k}}$, we can achieve

$$\min_{k \in [T]} \epsilon_k \le RG \frac{1 + \sum_{k=1}^{T} \frac{1}{k}}{2\sum_{k=1}^{T} \frac{1}{\sqrt{k}}} \le RG \frac{2 + \int_1^T \frac{1}{x} \mathrm{d}x}{4\int_1^{T+1} \frac{1}{\sqrt{x}} \mathrm{d}x} \le \frac{RG \log T}{2\sqrt{T}}. \tag{46}$$

Remark 1 The term $\log T$ in the bound can actually be removed by using the following simple fact. Given $c > 0$, $\mathbf{b} \in \mathbb{R}_+^d$, and D a positive *definite* matrix. Then

$$\min_{\mathbf{x} \in \mathbb{R}_+^d} \frac{c + \frac{1}{2}\mathbf{x}' D \mathbf{x}}{\mathbf{b}'\mathbf{x}} = \sqrt{\frac{2c}{\mathbf{b}' D^{-1} \mathbf{b}}}, \qquad \text{where the optimal } \mathbf{x} = \sqrt{\frac{2c}{\mathbf{b}' D^{-1} \mathbf{b}}} D^{-1}\mathbf{b}. \tag{47}$$

One can prove it by writing out the KKT condition for the equivalent convex problem (with a perspective function) $\inf_{\mathbf{x},u} \frac{1}{u}(c + \frac{1}{2}\mathbf{x}' D \mathbf{x})$, s.t. $\mathbf{x} \in \mathbb{R}_+^d$, $u > 0$, and $\mathbf{b}'\mathbf{x} = u$. Now apply this result to (45) with all $\eta_k = \frac{R}{G\sqrt{T}}$ ($k \in [T]$), then we get

$$\min_{k \in [T]} \epsilon_k \le \frac{RG}{\sqrt{T}}. \tag{48}$$

So to drive $\min_{k \in [T]} \epsilon_k$ below a threshold $\epsilon > 0$, it suffices to take T steps where

$$T \ge \frac{R^2 G^2}{\epsilon^2}. \tag{49}$$

Note the method requires that the horizon T be specified a priori, because the step size η_k needs this information. We next give a more intricate approach which does not require a pre-specified horizon.

Remark 2 The term $\log T$ in the bound can also be removed as follows. Here we redefine R^2 as the diameter square $\max_{x,y \in C} \|x - y\|_2^2$. Instead of telescoping over $k = 1, \ldots, T$, let us telescope from $k = T/2$ to T

(without loss of generality, let T be an even integer):

$$\|x_{T+1} - x^*\|_2^2 \le \left\|x_{T/2} - x^*\right\|_2^2 - 2\sum_{k=T/2}^{T} \eta_k \left(f(x_k) - f(x^*)\right) + \sum_{k=T/2}^{T} \eta_k^2 \|g_k\|_2^2 \tag{50}$$

$$\Longrightarrow \quad 2\sum_{k=T/2}^{T} \eta_k \left(f(x_k) - f(x^*)\right) \le R^2 + G^2 \sum_{k=T/2}^{T} \eta_k^2 \tag{51}$$

$$\Longrightarrow \quad \min_{k\in\{T/2,\ldots,T\}} \epsilon_k \le \frac{R^2 + G^2 \sum_{k=T/2}^{T} \eta_k^2}{2\sum_{k=T/2}^{T} \eta_k} \tag{52}$$

$$(\text{plug in } \eta_k = \frac{R}{G\sqrt{k}}) \quad = RG\frac{1+\sum_{k=T/2}^{T}\frac{1}{k}}{2\sum_{k=T/2}^{T}\frac{1}{\sqrt{k}}} \le RG\frac{1+\int_{T/2-1}^{T}\log x \mathrm{d}x}{4\int_{T/2}^{T+1}\sqrt{x}\mathrm{d}x} \le \frac{2RG}{\sqrt{T}}. \tag{53}$$

The trick is to exploit $\log T - \log(\frac{T}{2}-1) \approx \log 2$ in the numerator. In step (51), we bounded $\left\|x_{T/2} - x^*\right\|_2^2$ by R^2, because in general we cannot bound it by $\|x_1 - x^*\|_2^2$. In the sequel, we will simply write

$$\min_{k\in[T]} \epsilon_k \le \frac{RG}{\sqrt{T}}$$

ignoring the constants.

2.2 Rate of convergence for subgradient descent with mirror descent

The rate of convergence of subgradient descent often depends on R and G, which may depend unfortunately on the dimension of the problem. For example, suppose C is the simplex. Then $R \le \sqrt{2}$. If each coordinate of each gradient g_i is upper bounded by M, then G can be at most $M\sqrt{n}$, *i.e.* depends on the dimension of x.

We next see how this dependency can be removed by extending Euclidean distance to Bregman divergence. Clearly the steps 2 to 4 above can be easily extended by replacing $\|x_{k+1} - x^*\|_2^2$ with $\Delta_\psi(x^*, x_{k+1})$. So the only challenge left is to extend step 1. This is actually possible via Lemma 2.

We further assume ψ is σ strongly convex on C. In (33), consider $\eta_k(f(x_k) + \langle g_k, x - x_k\rangle)$ as the L in Lemma 2. Then

$$\eta_k \left(f(x_k) + \langle g_k, x^* - x_k\rangle\right) + \Delta_\psi(x^*, x_k) \ge \eta_k \left(f(x_k) + \langle g_k, x_{k+1} - x_k\rangle\right) + \Delta_\psi(x_{k+1}, x_k) \tag{54}$$
$$+ \Delta_\psi(x^*, x_{k+1}). \tag{55}$$

Canceling some terms can rearranging, we obtain

$$\Delta_\psi(x^*, x_{k+1}) \le \Delta_\psi(x^*, x_k) + \eta_k \langle g_k, x^* - x_{k+1}\rangle - \Delta_\psi(x_{k+1}, x_k) \tag{56}$$
$$= \Delta_\psi(x^*, x_k) + \eta_k \langle g_k, x^* - x_k\rangle + \eta_k \langle g_k, x_k - x_{k+1}\rangle - \Delta_\psi(x_{k+1}, x_k) \tag{57}$$
$$\le \Delta_\psi(x^*, x_k) - \eta_k \left(f(x_k) - f(x^*)\right) + \eta_k \langle g_k, x_k - x_{k+1}\rangle - \frac{\sigma}{2}\|x_k - x_{k+1}\|^2 \tag{58}$$
$$\le \Delta_\psi(x^*, x_k) - \eta_k \left(f(x_k) - f(x^*)\right) + \eta_k \|g_k\|_* \|x_k - x_{k+1}\| - \frac{\sigma}{2}\|x_k - x_{k+1}\|^2 \tag{59}$$
$$\le \Delta_\psi(x^*, x_k) - \eta_k \left(f(x_k) - f(x^*)\right) + \frac{\eta_k^2}{2\sigma}\|g_k\|_*^2 \tag{60}$$

Now compare with (42), we have successfully replaced $\|x_{k+1} - x^*\|_2^2$ with $\Delta_\psi(x^*, x_i)$. Again upper bound $\Delta_\psi(x^*, x_1)$ by R^2 and $\|g_k\|_*$ by G, and we obtain

$$\min_{k\in[T]} \epsilon_k \le \frac{RG}{\sqrt{\sigma T}}. \tag{61}$$

Note the norm on g_k is the dual norm. To see the advantage of mirror descent, suppose C is the n dimensional simplex, and we use KL-divergence for which ψ is 1 strongly convex with respect to the L_1 norm. The dual norm of the L_1 norm is the L_∞ norm. Then we can bound $\Delta_\psi(x^*, x_1)$ by using KL-divergence, and it is at most $\log n$ if we set $x_1 = \frac{1}{n}\mathbf{1}$ and x^* lies in the probability simplex. G can be upper bounded by M, and R by $\log n$. So with regard to the value of RG, mirror descent yields $M\log n$, which is smaller than that of subgradient descent by an order of $O(\sqrt{\frac{n}{\log n}})$. Note the saving of $\Theta(\sqrt{n})$ is from the norm of gradient (G) by replacing the L_2 norm by the L_∞ norm, at a slight cost of increasing R by $\log n$.

Remark 4 Note R^2 is an upper bound on $\Delta_\psi(x^*, x_1)$, rather than the real diameter $\max_{\mathbf{x},\mathbf{y}\in C} \Delta_\psi(x, y)$. This is important because for KL divergence defined on the probability simplex, the latter is actually infinity, while $\max_{x\in\Omega} \Delta_\psi(x, \frac{1}{n}\mathbf{1}) = \log n$.

2.3 Possibilities for accelerated rates

When the objective function has additional properties, the rates can be significantly improved. Here we see two examples.

Acceleration 1: f is strongly convex. We say f is strongly convex with respect to another convex function ψ with modulus λ if

$$f(x) \geq f(y) + \langle g, x - y\rangle + \lambda\Delta_\psi(x, y) \qquad \forall\, g \in \partial f(y). \tag{62}$$

Note we do not assume f is differentiable. Now in the step from (57) to (58), we can plug in the definition of strong convexity:

$$\begin{aligned}
\Delta_\psi(x^*, x_{k+1}) &= \ldots + \eta_k \langle g_k, x^* - x_k\rangle + \ldots \qquad\qquad \text{(copy of (57))} & (63)\\
&\leq \ldots - \eta_k \left(f(x_k) - f(x^*) + \lambda\Delta_\psi(x^*, x_k)\right) + \ldots & (64)\\
&\leq \ldots & (65)\\
&\leq (1 - \lambda\eta_k)\Delta_\psi(x^*, x_k) - \eta_k \left(f(x_k) - f(x^*)\right) + \frac{\eta_k^2}{2\sigma}\left\|g_k\right\|_*^2 & (66)
\end{aligned}$$

Denote $\delta_k = \Delta_\psi(x^*, x_k)$. Set $\eta_k = \frac{1}{\lambda k}$. Then

$$\delta_{k+1} \leq \frac{k-1}{k}\delta_k - \frac{1}{\lambda k}\epsilon_k + \frac{G^2}{2\sigma\lambda^2 k^2} \quad \Longrightarrow \quad k\delta_{k+1} \leq (k-1)\delta_k - \frac{1}{\lambda}\epsilon_k + \frac{G^2}{2\sigma\lambda^2 k} \tag{67}$$

Now telescope (sum up both sides from $k = 1$ to T)

$$T\delta_{T+1} \leq -\frac{1}{\lambda}\sum_{k=1}^T \epsilon_k + \frac{G^2}{2\sigma\lambda^2}\sum_{k=1}^T \frac{1}{k} \quad \Longrightarrow \quad \min_{i\in[T]} \epsilon_k \leq \frac{G^2}{2\sigma\lambda}\frac{1}{T}\sum_{k=1}^T \frac{1}{k} \leq \frac{G^2}{2\sigma\lambda}\frac{O(\log T)}{T}. \tag{68}$$

Acceleration 2: f has Lipschitz continuous gradient. If the gradient of f is Lipschitz continuous, there exists $L > 0$ such that

$$\left\|\nabla f(x) - \nabla f(y)\right\|_* \leq L\left\|x - y\right\|, \qquad \forall\, x, y. \tag{69}$$

Sometimes we just directly say f is smooth. It is also known that this is equivalent to

$$f(x) \leq f(y) + \langle\nabla f(y), x - y\rangle + \frac{L}{2}\left\|x - y\right\|^2. \tag{70}$$

We bound the $\langle g_k, x^* - x_{k+1}\rangle$ term in (56) as follows

$$\begin{aligned}
\langle g_k, x^* - x_{k+1}\rangle &= \langle g_k, x^* - x_k\rangle + \langle g_k, x_k - x_{k+1}\rangle & (71)\\
&\leq f(x^*) - f(x_k) + f(x_k) - f(x_{k+1}) + \frac{L}{2}\left\|x_k - x_{k+1}\right\|^2 & (72)\\
&= f(x^*) - f(x_{k+1}) + \frac{L}{2}\left\|x_k - x_{k+1}\right\|^2. & (73)
\end{aligned}$$

Plug into (56), we get

$$\Delta_\psi(x^*, x_{k+1}) \leq \Delta_\psi(x^*, x_k) + \eta_k\left(f(x^*) - f(x_{k+1}) + \frac{L}{2}\left\|x_k - x_{k+1}\right\|^2\right) - \frac{\sigma}{2}\left\|x_k - x_{k+1}\right\|^2. \tag{74}$$

Set $\eta_k = \frac{\sigma}{L}$, we get

$$\Delta_\psi(x^*, x_{k+1}) \leq \Delta_\psi(x^*, x_k) - \frac{\sigma}{L}\left(f(x_{k+1}) - f(x^*)\right). \tag{75}$$

Telescope we get

$$\min_{k\in\{2,\ldots,T+1\}} f(x_k) - f(x^*) \;\leq\; \frac{L\Delta(x^*, x_1)}{\sigma T} \leq \frac{LR^2}{\sigma T}. \tag{76}$$

This gives $O(\frac{1}{T})$ convergence rate. But if we are smarter, like Nesterov, the rate can be improved to $O(\frac{1}{T^2})$. We will not go into the details but the algorithm and proof are again based on Lemma 2. This is often called accelerated proximal gradient method.

2.4 Composite Objective

Suppose the objective function is $h(x) = f(x) + r(x)$, where f is smooth and $r(x)$ is simple, like $\|x\|_1$. If we directly apply the above rates for optimizing h, we get $O(\frac{1}{\sqrt{T}})$ rate of convergence because h is not smooth. It will be nice if we can enjoy the $O(\frac{1}{T})$ rate as in smooth optimization. Fortunately this is possible thanks to the simplicity of $r(x)$, and we only need to extend the proximal operator (33) as follows:

$$x_{k+1} = \operatorname*{argmin}_{x \in C} \left\{ f(x_k) + \langle g_k, x - x_k \rangle + r(x) + \frac{1}{\eta_k} \Delta_\psi(x, x_k) \right\} \tag{77}$$

$$= \operatorname*{argmin}_{x \in C} \left\{ \eta_k f(x_k) + \eta_k \langle g_k, x - x_k \rangle + \eta_k r(x) + \Delta_\psi(x, x_k) \right\}. \tag{78}$$

Here we use a first-order Taylor approximation of f around x_k, but keep $r(x)$ exact. Assuming this proximal operator can be computed efficiently, then we can show all the above rates carry over. We here only show the case of general f (not necessarily strongly convex or has Lipschitz continuous gradient), and leave the rest two cases as an exercise. In fact we can again achieve $O(\frac{1}{T^2})$ rate when f has Lipschitz continuous gradient.

Consider $\eta_k(f(x_k) + \langle g_k, x - x_k \rangle + r(x))$ as the L in Lemma 2. Then

$$\eta_k \left(f(x_k) + \langle g_k, x^* - x_k \rangle + r(x^*) \right) + \Delta_\psi(x^*, x_k) \tag{79}$$

$$\geq \quad \eta_k \left(f(x_k) + \langle g_k, x_{k+1} - x_k \rangle + r(x_{k+1}) \right) + \Delta_\psi(x_{k+1}, x_k) + \Delta_\psi(x^*, x_{k+1}). \tag{80}$$

Following exactly the derivations from (56) to (60), we obtain

$$\Delta_\psi(x^*, x_{k+1}) \leq \Delta_\psi(x^*, x_k) + \eta_k \langle g_k, x^* - x_{k+1} \rangle + \eta_k (r(x^*) - r(x_{k+1})) - \Delta_\psi(x_{k+1}, x_k) \tag{81}$$

$$\leq \ldots \tag{82}$$

$$\leq \Delta_\psi(x^*, x_k) - \eta_k \left(f(x_k) + r(x_{k+1}) - f(x^*) - r(x^*) \right) + \frac{\eta_k^2}{2\sigma} \|g_k\|_*^2. \tag{83}$$

This is almost the same as (60), except that we want to have $r(x_k)$ here, not $r(x_{k+1})$. Fortunately this is not a problem as long as we use a slightly different way of telescoping. Denote $\delta_k = \Delta_\psi(x^*, x_k)$ and then

$$f(x_k) + r(x_{k+1}) - f(x^*) - r(x^*) \leq \frac{1}{\eta_k}(\delta_k - \delta_{k+1}) + \frac{\eta_k}{2\sigma} \|g_k\|_*^2. \tag{84}$$

Summing up from $k = 1$ to T we obtain

$$r(x_{T+1}) - r(x_1) + \sum_{k=1}^{T} (h(x_k) - h(x^*)) \leq \frac{\delta_1}{\eta_1} + \sum_{k=2}^{T} \delta_k \left(\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}} \right) - \frac{\delta_{T+1}}{\eta_T} + \frac{G^2}{2\sigma} \sum_{k=1}^{T} \eta_k \tag{85}$$

$$\leq R^2 \left(\frac{1}{\eta_1} + \sum_{k=2}^{T} \left(\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}} \right) \right) + \frac{G^2}{2\sigma} \sum_{k=1}^{T} \eta_k \tag{86}$$

$$= \frac{R^2}{\eta_T} + \frac{G^2}{2\sigma} \sum_{k=1}^{T} \eta_k. \tag{87}$$

Suppose we choose $x_1 = \operatorname{argmin}_x r(x)$, which ensures $r(x_{T+1}) - r(x_1) \geq 0$. Setting $\eta_k = \frac{R}{G}\sqrt{\frac{\sigma}{k}}$, we get

$$\sum_{k=1}^{T} (h(x_k) - h(x^*)) \leq \frac{RG}{\sqrt{\sigma}} \left(\sqrt{T} + \frac{1}{2} \sum_{k=1}^{T} \frac{1}{\sqrt{k}} \right) = \frac{RG}{\sqrt{\sigma}} O(\sqrt{T}). \tag{88}$$

Therefore $\min_{k \in [T]} \{h(x_k) - h(x^*)\}$ decays at the rate of $O(\frac{RG}{\sqrt{\sigma T}})$.

3 Online and Stochastic Learning

The protocol of online learning is shown in Algorithm 1. The player's goal of online learning is to minimize the regret, the minimal possible loss $\sum_k f_k(x)$ over all possible x:

$$\text{Regret} = \sum_{k=1}^{T} f_k(x_k) - \min_x \sum_{k=1}^{T} f_k(x). \tag{89}$$

Note there is no assumption made on how the rival picks f_k, and it can adversarial. After obtaining f_k at iteration k, let us update the model into x_{k+1} by using the mirror descent rule on function f_k only:

$$x_{k+1} = \operatorname*{argmin}_{x \in C} \left\{ f_k(x_k) + \langle g_k, x - x_k \rangle + \frac{1}{\eta_k} \Delta_\psi(x, x_k) \right\}, \quad \text{where} \quad g_k \in \partial f_k(x_k). \tag{90}$$

Algorithm 1: Protocol of online learning

1 The player initializes a model x_1.
2 **for** $k = 1, 2, \ldots$ **do**
3 The player proposes a model x_k.
4 The rival picks a function f_k.
5 The player suffers a loss $f_k(x_k)$.
6 The player gets access to f_k and use it to update its model to x_{k+1}.

Then it is easy to derive the regret bound. Using f_k in step (60), we have

$$f_k(x_k) - f_k(x^*) \leq \frac{1}{\eta_k}(\Delta_\psi(x^*, x_k) - \Delta_\psi(x^*, x_{k+1})) + \frac{\eta_k}{2\sigma} \|g_k\|_*^2 . \tag{91}$$

Summing up from $k = 1$ to n and using the same process as in (85) to (88), we get

$$\sum_{k=1}^{T}(f_k(x_k) - f_k(x^*)) \leq \frac{RG}{\sqrt{\sigma}} O(\sqrt{T}). \tag{92}$$

So the regret grows in the order of $O(\sqrt{T})$.

f is strongly convex. Exactly use (66) with f_k in place of f, and we can derive the $O(\log T)$ regret bound immediately.

f has Lipschitz continuous gradient. The result in (75) can NOT be extended to the online setting because if we replace f by f_k we will get $f_k(x_{k+1}) - f_k(x^*)$ on the right-hand side. Telescoping will not give a regret bound. In fact, it is known that in the online setting, having a Lipschitz continuous gradient itself cannot reduce the regret bound from $O(\sqrt{T})$ (as in nonsmooth objective) to $O(\log T)$.

Composite objective. In the online setting, both the player and the rival know $r(x)$, and the rival changes $f_k(x)$ at each iteration. The loss incurred at each iteration is $h_k(x_k) = f_k(x_k) + r(x_k)$. The update rule is

$$x_{k+1} = \operatorname*{argmin}_{x \in C} \left\{ f_k(x_k) + \langle g_k, x - x_k \rangle + r(x) + \frac{1}{\eta_k} \Delta_\psi(x, x_k) \right\}, \quad \text{where} \quad g_k \in \partial f_k(x_k). \tag{93}$$

Note in this setting, (84) becomes

$$f_k(x_k) + r(x_{k+1}) - f_k(x^*) - r(x^*) \leq \frac{1}{\eta_k}(\delta_k - \delta_{k+1}) + \frac{\eta_k}{2\sigma} \|g_k\|_*^2 . \tag{94}$$

Although we have $r(x_{k+1})$ here rather than $r(x_k)$, it is fine because r does not change through iterations. Choosing $x_1 = \operatorname{argmin}_x r(x)$ and telescoping in the same way as from (85) to (88), we immediately obtain

$$\sum_{k=1}^{T}(h_k(x_k) - h_k(x^*)) \leq \frac{G}{\sqrt{\sigma}} O(\sqrt{T}). \tag{95}$$

So the regret grows at $O(\sqrt{T})$.

When f_k are strongly convex, we can get $O(\log T)$ regret for the composite case. But as expected, having Lipschitz continuity of ∇f_k alone cannot reduce the regret from $O(\sqrt{T})$ to $O(\log T)$.

3.1 Stochastic optimization

Let us consider optimizing a function which takes a form of expectation

$$\min_x \; F(x) := \mathop{\mathbb{E}}_{\omega \sim p} [f(x; \omega)], \tag{96}$$

where p is a distribution of ω. This subsumes a lot of machine learning models. For example, the SVM objective is

$$F(x) = \frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - c_i \langle a_i, x \rangle\} + \frac{\lambda}{2} \|x\|^2 . \tag{97}$$

Algorithm 2: Protocol of online learning

```
1 The player initializes a model x_1.
2 for k = 1, 2, ... do
3 |  The player proposes a model x_k.
4 |  The rival randomly draws a ω_k from p, which defines a function f_k(x) := f(x; ω_k).
5 |  The player suffers a loss f_k(x_k).
6 |  The player gets access to f_k and use it to update its model to x_{k+1} by, e.g., mirror descent (90).
```

It can be interpreted as (96) where ω is uniformly distributed in $\{1, 2, \ldots, m\}$ (*i.e.* $p(\omega = i) = \frac{1}{m}$), and

$$f(x; i) = \max\{0, 1 - c_i \langle a_i, x \rangle\} + \frac{\lambda}{2} \|x\|^2 . \tag{98}$$

When m is large, it can be costly to calculate F and its subgradient. So a simple idea is to base the updates on a single randomly chosen data point. It can be considered as a special case of online learning in Algorithm 1, where the rival in step 4 now randomly picks f_k as $f(x; \omega_k)$ with ω_k being drawn independently from p. Ideally we hope that by using the mirror descent updates, x_k will gradually approach the minimizer of $F(x)$. Intuitively this is quite reasonable, and by using f_k we can compute an unbiased estimate of $F(x_k)$ and a subgradient of $F(x_k)$ (because ω_k are sampled iid from p). This is a particular case of stochastic optimization, and we recap it in Algorithm 2.

In fact, the method is valid in a more general setting. For simplicity, let us just say the rival plays ω_k at iteration k. Then an online learning algorithm $\mathcal{A}$ is simply a deterministic mapping from an ordered set $\{\omega_1, \ldots, \omega_k\}$ to x_{k+1}. Denote as $\mathcal{A}(\omega_0)$ the initial model x_1. Then the following theorem is the key for online to batch conversion.

Theorem 3 *Suppose an online learning algorithm $\mathcal{A}$ has regret bound R_k after running Algorithm 1 for k iterations. Suppose $\omega_1, \ldots, \omega_{T+1}$ are drawn iid from p. Define $\hat{x} = \mathcal{A}(\omega_{j+1}, \ldots, \omega_T)$ where j is drawn uniformly random from $\{0, \ldots, T\}$. Then*

$$\mathbb{E}[F(\hat{x})] - \min_x F(x) \le \frac{R_{T+1}}{T+1}, \tag{99}$$

where the expectation is with respect to the randomness of $\omega_1, \ldots, \omega_T$, and j.

Similarly we can have high probability bounds, which can be stated in the form like (not exactly true)

$$F(\hat{x}) - \min_x F(x) \le \frac{R_{T+1}}{T+1} \log \frac{1}{\delta} \tag{100}$$

with probability $1 - \delta$, where the probability is with respect to the randomness of $\omega_1, \ldots, \omega_T$, and j.

Proof of Theorem 3.

$$\mathbb{E}[F(\hat{x})] = \mathop{\mathbb{E}}_{j, \omega_1, \ldots, \omega_{T+1}} [f(\hat{x}; \omega_{T+1})] = \mathop{\mathbb{E}}_{j, \omega_1, \ldots, \omega_{T+1}} [f(\mathcal{A}(\omega_{j+1}, \ldots, \omega_T); \omega_{T+1})] \tag{101}$$

$$= \mathop{\mathbb{E}}_{\omega_1, \ldots, \omega_{T+1}} \left[\frac{1}{T+1} \sum_{j=0}^{T} f(\mathcal{A}(\omega_{j+1}, \ldots, \omega_T); \omega_{T+1}) \right] \quad \text{(as } j \text{ is drawn uniformly random)} \tag{102}$$

$$= \frac{1}{T+1} \mathop{\mathbb{E}}_{\omega_1, \ldots, \omega_{T+1}} \left[\sum_{j=0}^{T} f(\mathcal{A}(\omega_1, \ldots, \omega_{T-j}); \omega_{T+1-j}) \right] \quad \text{(shift iteration index by iid of } w_i\text{)} \tag{103}$$

$$= \frac{1}{T+1} \mathop{\mathbb{E}}_{\omega_1, \ldots, \omega_{T+1}} \left[\sum_{s=1}^{T+1} f(\mathcal{A}(\omega_1, \ldots, \omega_{s-1}); \omega_s) \right] \quad \text{(change of variable } s = T - j + 1\text{)} \tag{104}$$

$$\le \frac{1}{T+1} \mathop{\mathbb{E}}_{\omega_1, \ldots, \omega_{T+1}} \left[\min_x \sum_{s=1}^{T+1} f(x; \omega_s) + R_{T+1} \right] \quad \text{(apply regret bound)} \tag{105}$$

$$\le \min_x \mathop{\mathbb{E}}_{\omega} [f(x; \omega)] + \frac{R_{T+1}}{T+1} \quad \text{(expectation of min is smaller than min of expectation)} \tag{106}$$

$$= \min_x F(x) + \frac{R_{T+1}}{T+1}. \tag{107}$$