convex notes

September 26, 2023

1 Gradient descent

Definition 1 (β-smooth)**.** *A function f is (β, q)-smooth if the gradient ∇f is β-Lipschitz in the dual norm:*

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\| . \tag{1}$$

Lemma 2 ((1) $\implies$ (2))**.** *Let f be a β-smooth function on $\mathbb{R}^n$. Then, for any $x, y \in \mathbb{R}^n$,*

$$f(x) - f(y) - \nabla f(y)^T (x - y) \leq \frac{\beta}{2} \|x - y\|^2 . \tag{2}$$

Proof. We first represent $f(x) - f(y)$ as an integral. Let $g(t) = f(y + t(x - y))$, so that $g'(t) = \nabla f(y + t(x - y))^T (x - y)$, since it's the rate of change of f at point $y + t(x - y)$ in the direction $x - y$. By fundamental theorem of calculus,

$$f(x) - f(y) = g(1) - g(0) = \int_0^1 g'(t) dt = \int_0^1 \nabla f(y + t(x - y))^T (x - y) dt.$$

We apply Cauchy-Schwarz and then β-smoothness:

$$\begin{aligned}
f(x) - f(y) - \nabla f(y)^T (x - y) &= \int_0^1 \nabla f(y + t(x - y))^T (x - y) dt - \nabla f(y)^T (x - y) \\
&= \int_0^1 \big(\nabla f(y + t(x - y)) - \nabla f(y)\big)^T (x - y) dt \\
&\leq \int_0^1 \|\nabla f(y + t(x - y)) - \nabla f(y)\| \cdot \|x - y\| \, dt \\
&\leq \int_0^1 \beta t \|x - y\|^2 \, dt \\
&= \frac{\beta}{2} \|x - y\|^2 .
\end{aligned}$$

□

Therefore, if f is both convex and β-smooth, then

$$0 \leq f(x) - f(y) - \nabla f(y)^T (x - y) \leq \frac{\beta}{2} \|x - y\|^2 . \tag{3}$$

In fact, Equations (1) and (2) are equivalent for convex functions, so we could have defined β-smooth using either equation.

Lemma 3 ((2) $\implies$ (1))**.** *Let f be a convex function satisfying (2). Then, f is β-smooth.*

Proof. Let

$$z = y - \frac{1}{\beta}(\nabla f(y) - \nabla f(x)).$$

Then,

$$\begin{aligned}
f(x) - f(y) &= f(x) - f(z) + f(z) - f(y) \\
&\overset{\text{convex},(2)}{\leq} \nabla f(x)^T(x-z) + \nabla f(y)^T(z-y) + \frac{\beta}{2}\|z-y\|^2 \\
&= \nabla f(x)^T(x-y) + (\nabla f(x) - \nabla f(y))^T(y-z) + \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 \\
&= \nabla f(x)^T(x-y) + (\nabla f(x) - \nabla f(y))^T\left(\frac{1}{\beta}(\nabla f(y) - \nabla f(x))\right) + \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 \\
&= \nabla f(x)^T(x-y) - \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2.
\end{aligned}$$

Rearranging,

$$\frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 \leq f(y) - f(x) - \nabla f(x)^T(y-x) \overset{(2)}{\leq} \frac{\beta}{2}\|x-y\|^2.$$

Taking the square root finishes the proof. ∎

Consider a gradient step $y = x - \frac{1}{\beta}\nabla f(x)$. From (2), we get

$$f\left(x - \frac{1}{\beta}\nabla f(x)\right) - f(x) + \nabla f(x)^T\left(\frac{1}{\beta}\nabla f(x)\right) = f(y) - f(x) - \nabla f(x)^T(y-x) \leq \frac{\beta}{2}\|x-y\|^2 = \frac{\beta}{2}\left\|\frac{1}{\beta}\nabla f(x)\right\|^2$$

$$\iff f\left(x - \frac{1}{\beta}\nabla f(x)\right) - f(x) \leq -\frac{1}{2\beta}\|\nabla f(x)\|^2. \tag{4}$$

Basically, the step size $\eta = \frac{1}{\beta}$ is chosen small enough that the linear (in the step size) term $-\nabla f(x)^T(y-x)$ dominates the quadratic term $\frac{\beta}{2}\left\|\frac{1}{\beta}\nabla f(x)\right\|^2$.

Theorem 4. *Let f be convex and β-smooth. Then, gradient descent with $\eta = \frac{1}{\beta}$ satisfies*

$$f(x_t) - f(x^*) \leq \frac{2\beta\|x_1 - x^*\|^2}{t-1}.$$

Proof. By (4), we have

$$f(x_{x+1}) - f(x_s) \leq -\frac{1}{2\beta}\|\nabla f(x_s)\|^2.$$

Let $\delta_s = f(x_s) - f(x^*)$ be how close the current point is to optimal. Rewriting,

$$\delta_{s+1} \leq \delta_s - \frac{1}{2\beta}\|\nabla f(x_s)\|^2.$$

Also, by convexity,

$$\delta_{s+1} = f(x_s) - f(x^*) \leq \nabla f(x_s)^T(x_s - x^*) \leq \|x_s - x^*\| \cdot \|\nabla f(x_s)\|,$$

which, intuitively, means that $\nabla f(x_s)$ should decrease at least as much as if the function were a straight line between x_s and x^* (because f is convex). We will prove that $\|x_s - x^*\|$ is decreasing with s; assuming this, we obtain

$$\delta_{s+1} \leq \|x_s - x^*\| \cdot \|\nabla f(x_s)\| \leq \|x_1 - x^*\| \cdot \|\nabla f(x_s)\|$$

$$\implies \delta_{s+1} \le \delta_s - \frac{1}{2\beta}\left\|\nabla f(x_s)\right\|^2 \le \delta_s - \frac{1}{2\beta}\left(\frac{\delta_{s+1}}{\|x_1 - x^*\|}\right)^2 = \delta_s - \frac{1}{2\beta\left\|x_1 - x^*\right\|^2}\delta_s^2.$$

To finish the theorem, iterate and prove by induction.

To show that $\|x_s - x^*\|$ decreases, we use (2) twice, with x and y exchanged, to obtain

$$(\nabla f(x) - \nabla f(y))^T(x - y) \ge \frac{1}{\beta}\left\|\nabla f(x) - \nabla f(y)\right\|^2.$$

Plugging in $x \leftarrow x_s$ and $y \leftarrow x^*$ and using that $\nabla f(x^*) = 0$,

$$\nabla f(x_s)^T(x_s - x^*) \ge \frac{1}{\beta}\left\|\nabla f(x)\right\|^2.$$

Therefore,

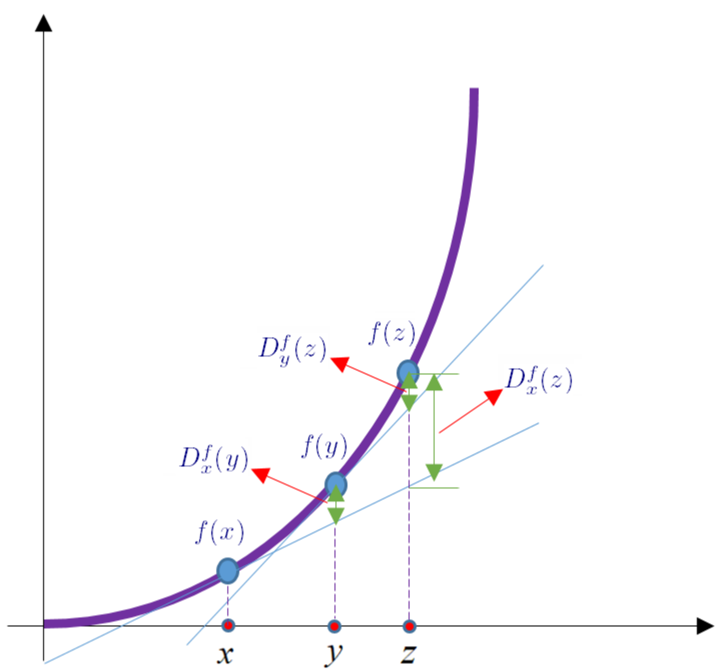
$$\begin{aligned}
\|x_{s+1} - x^*\|^2 = \left\|x_s - \frac{1}{\beta}\nabla f(x_s) - x^*\right\|^2 &= \|x_s - x^*\|^2 - \frac{2}{\beta}\nabla f(x_s)^T(x_s - x^*) + \frac{1}{\beta^2}\left\|\nabla f(x_s)\right\|^2 \\
&\le \|x_s - x^*\|^2 - \frac{2}{\beta^2}\left\|\nabla f(x_s)\right\|^2 + \frac{1}{\beta^2}\left\|\nabla f(x_s)\right\|^2 \\
&= \|x_s - x^*\|^2 - \frac{1}{\beta^2}\left\|\nabla f(x_s)\right\|^2 \\
&\le \|x_s - x^*\|^2.
\end{aligned}$$

□

2 Mirror descent

Definition 5 (Bregman divergence)**.** $V_x^r(y) = r(y) - r(x) - \langle\nabla r(x), y - x\rangle$

Lemma 6 (Three-point equality)**.** $V_x^r(y) + V_y^r(z) = V_x^r(z) + \langle\nabla r(x) - \nabla r(y), z - y\rangle$



Proof. Expanding the Bregman divergence terms, the $r(x), r(y), r(z)$ terms match. For the inner product terms, we have $-\langle \nabla r(x), y-x\rangle - \langle \nabla r(y), z-y\rangle$ on the left and $-\langle \nabla r(x), z-x\rangle + \langle \nabla r(x) - \nabla r(y), z-y\rangle$ on the right. These terms clearly match up. $\square$

Lemma 7 (Three-point inequality)**.** *Let f be a smooth function and let $y = \arg\min\{\langle \nabla f(x), y\rangle + V_x^r(y)\}$. Then, for any z in the (convex) domain,*

$$\langle \nabla f(x), y-z\rangle \leq V_x^r(z) - V_x^r(y) - V_y^r(z)$$

Proof. Re-arranging the three-point equality,

$$\langle \nabla r(x) - \nabla r(y), y-z\rangle = V_x^r(z) - V_x^r(y) - V_y^r(z).$$

It remains to show that

$$\langle \nabla f(x) - \nabla r(x) + \nabla r(y), y-z\rangle \leq 0.$$

Suppose first that the domain is everything. Then, the minimizer y satisfies

$$0 = \nabla_y(\langle \nabla f(x), y\rangle + V_x^r(y)) = \nabla_y\big(\langle \nabla f(x), y\rangle + r(y) - r(x) - \langle \nabla r(x), y-x\rangle\big) = \nabla f(x) + \nabla r(y) - \nabla r(x),$$

so the inequality above is actually an equality. More generally, we must have $\langle \nabla f(x) + \nabla r(y) - \nabla r(x), z-y\rangle \geq 0$ since moving the solution from the minimizer y in the direction of z can only increase the function value. $\square$

Mirror descent. Start with an arbitrary x_0. For each iteration $t \in [T]$, let $x_{t+1} = \arg\min\{\langle \nabla f(x_t), y\rangle + V_{x_t}^r(y)\}$.

Lemma 8. *Assume that*

1. *$V_x^r(y)$ is 1-strongly convex for all x, y in the primal norm: $V_x^r(y) \geq \frac{1}{2}\|x-y\|^2$*
2. *$V_x^r(y) \leq R$ for all x, y in the domain*
3. *$\|\nabla f(x)\|_* \leq L$ for all x in the domain.*

Then, running mirror descent for T iterations gives

$$\sum_{t=0}^{T-1} \langle \nabla f(x_t), x_t - x^*\rangle \leq R + \frac{L^2 T}{2}.$$

Proof. By the three-point inequality,

$$\langle \nabla f(x_t), x_{t+1} - x^*\rangle \leq V_{x_t}^r(x^*) - V_{x_t}^r(x_{t+1}) - V_{x_{t+1}}^r(x^*) \leq V_{x_t}^r(x^*) - \frac{1}{2}\|x_t - x_{t+1}\|^2 - V_{x_{t+1}}^r(x^*).$$

Adding $\langle \nabla f(x_t), x_t - x_{t+1}\rangle$ to both sides, and then applying Cauchy-Schwarz and then AM-GM on

$$\langle \nabla f(x_t), x_t - x_{t+1}\rangle \leq \|\nabla f(x_t)\|_* \cdot \|x_t - x_{t+1}\| \leq \frac{1}{2}\|\nabla f(x_t)\|_*^2 + \frac{1}{2}\|x_t - x_{t+1}\|^2,$$

we obtain

$$\langle \nabla f(x_t), x_t - x^*\rangle \leq V_{x_t}^r(x^*) - V_{x_{t+1}}^r(x^*) + \frac{1}{2}\|\nabla f(x_t)\|_*^2 \leq V_{x_t}^r(x^*) - V_{x_{t+1}}^r(x^*) + \frac{L^2}{2}.$$

Summing over all $t \in [0, T-1]$, the terms $V_{x_t}^r(x^*) - V_{x_{t+1}}^r(x^*)$ telescope, and we obtain

$$\sum_{t=0}^{T-1} \langle \nabla f(x_t), x_t - x^*\rangle \leq V_{x_0}^r(x^*) - V_{x_T}^r(x^*) + \frac{L^2 T}{2}.$$

The lemma follows from $0 \leq V_x^r(y) \leq R$ for all x, y by assumption. $\square$

Suppose that we instead set $x_{t+1} = \arg\min\{\langle \eta \nabla f(x_t), y\rangle + V^r_{x_t}(y)\}$ for some parameter $\eta > 0$. We can essentially replace $f(x)$ with $\eta f(x)$ in the lemma above, which gives the guarantee

$$\eta \sum_{t=0}^{T-1} \langle \nabla f(x_t), x_t - x^* \rangle \leq R + \frac{(\eta L)^2 T}{2}.$$

The average regret is $\frac{1}{\eta T}(R + \eta^2 L^2 T/2) = R/(\eta T) + \eta L^2/2$, and optimizing η gives $O(\sqrt{RL^2/T})$, which is decreasing in T. Finally, setting $T = O(RL^2/\epsilon^2)$ gives average regret ϵ. Note that this coincides with multiplicative weights intuition: the dependency on ϵ is $1/\epsilon^2$, the dependency on the diameter R is linear (usually $O(\log n)$ for multiplicative weights), and the dependency on the width L is quadratic.

3 old mirror descent notes

Define the Bregman divergence

$$D_f(x, y) := f(x) - f(y) - \langle \nabla f(y), x - y \rangle,$$

which is always nonnegative when f is convex.

Mirror descent begins with x_0 arbitrarily, and then sets

- $y_i \leftarrow \nabla f(x_i)$ (mirror map to dual)
- $y_{i+1} \leftarrow y_i - \eta \nabla f(x_i)$ (take gradient step in dual)
- $x_{i+1} \leftarrow \nabla f^*(y_{i+1})$, i.e., select x_{i+1} s.t. $y_{i+1} = \nabla f(x_{i+1})$ (mirror map back)

Let $\overline{x} := \frac{1}{T}\sum_{i=0}^{T-1} x_i$ be the average. We want to show that

$$f(\overline{x}) - f(x) \overset{!}{\leq} \epsilon.$$

We begin with

$$f(\overline{x}) = f\left(\frac{1}{T}\sum_{i=0}^{T-1} x_i\right) \leq \frac{1}{T}\sum_{i=0}^{T-1} f(x_i),$$

so

$$f(\overline{x}) - f(x) \leq \frac{1}{T}\sum_{i=0}^{T-1} f(x_i) - f(x) \leq \frac{1}{T}\sum_{i=0}^{T-1} \big(f(x_i) - f(x)\big) \leq \frac{1}{T}\sum_{i=0}^{T-1} \langle \nabla f(x_i), x_i - x \rangle.$$

To see the last inequality, simply flip the sign of the convexity inequality $f(x) - f(x_i) \geq \langle \nabla f(x_i), x - x_i \rangle$.

Since the gradient steps in the dual are $\eta \nabla f(x_i)$, we can rewrite it as

$$f(\overline{x}) - f(x) \leq \frac{1}{T}\sum_{i=0}^{T-1} \langle \nabla f(x_i), x_i - x \rangle = \frac{1}{T}\sum_{i=0}^{T-1} \langle \frac{1}{\eta}(y_i - y_{i+1}), x_i - x \rangle \frac{1}{T}\sum_{i=0}^{T-1} \langle \frac{1}{\eta}(\nabla f(x_i) - \nabla f(x_{i+1})), x_i - x \rangle.$$

Let's now write

$$\langle \nabla f(x_i) - \nabla f(x_{i+1}), x_i - x \rangle$$

in terms of Bregman divergences. First, to capture the $\nabla f(x_i)$ and $\nabla f(x_{i+1})$ factors, we use

$$D_f(x, x_i) = f(x) - f(x_i) - \langle \nabla f(x_i), x - x_i \rangle,$$
$$D_f(x, x_{i+1}) = f(x) - f(x_{i+1}) - \langle \nabla f(x_{i+1}), x - x_{i+1} \rangle,$$

so we want the factors

$$
\begin{aligned}
+D_f(x, x_i) - D_f(x, x_{i+1}) &= -f(x_i) + f(x_{i+1}) - \langle \nabla f(x_i), x - x_i \rangle + \langle \nabla f(x_{i+1}), x - x_{i+1} \rangle \\
&= -f(x_i) + f(x_{i+1}) + \langle \nabla f(x_i), x_i - x \rangle - \langle \nabla f(x_{i+1}), x_{i+1} \rangle + \langle \nabla f(x_{i+1}), x \rangle,
\end{aligned}
$$

so we need to correct it by adding a

$$D_f(x_{i+1}, x_i) = f(x_{i+1}) - f(x_i) - \langle \nabla f(x_i), x_{i+1} - x_i \rangle$$

factor, which is exactly what we need. In other words,

$$\langle \nabla f(x_i) - \nabla f(x_{i+1}), x_i - x \rangle = D_f(x, x_i) - D_f(x, x_{i+1}) + D_f(x_{i+1}, x_i).$$

The first two terms are nice: they telescope once we sum over all $\langle \nabla f(x_i) - \nabla f(x_{i+1}), x_i - x \rangle$. The last term is what we want to show is small. We want it on the order of $\eta^{1+\delta}$, since we pay a factor $1/\eta$ at the end.