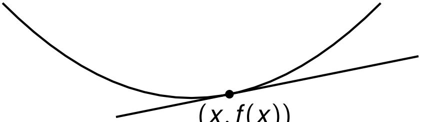
Convex optimization

Slides: Thanks to Di Wang.

$$\min_{x \in Q} f(x)$$

$$f(x) - f(y) \leq \langle \nabla f(x), x - y \rangle$$

Q: feasible space, convex.



Convex optimization

Slides: Thanks to Di Wang.

$$\min_{x \in Q} f(x)$$

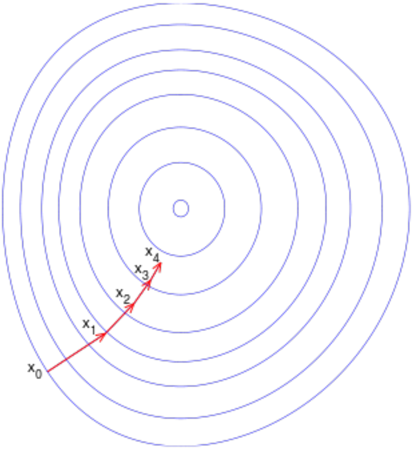
$$f(x) - f(y) \leq \langle \nabla f(x), x - y \rangle$$

Q: feasible space, convex.

First-order Iterative Methods

- Query $x \in Q$, update using $\nabla f(x)$
- Low per-iteration cost, $\mathrm{poly}(\frac{1}{\varepsilon})$ convergence.
- Methods of choice in large-scale regime.

Gradient Descent



- Moves in down-hill direction.
- Improve objective function value each iteration.
- Output final point.

Gradient Descent

L-Lipschitz continuous

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \quad \forall x, y \in Q$$

- Global linear lower bound and quadratic upper bound:

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

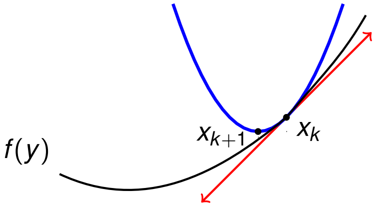
Gradient Descent

L-Lipschitz continuous

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \quad \forall x, y \in Q$$

- Global linear lower bound and quadratic upper bound:

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



$f(x)$ rises faster when going up,
and falls slower when going down.

Gradient Descent

L-Lipschitz continuous

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \quad \forall x, y \in Q$$

- Global linear lower bound and quadratic upper bound:

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

- Minimize using quadratic bound

$$x_{k+1} = \mathrm{Grad}(x_k) = \operatorname*{argmin}_{x \in Q} \{ \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2}\|x - x_k\|^2 \}$$

If $Q = \mathbb{R}^n$ and ℓ_2-norm, $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$.

Gradient Descent

L-Lipschitz continuous

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \quad \forall x, y \in Q$$

- Global linear lower bound and quadratic upper bound:

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

- Minimize using quadratic bound

$$x_{k+1} = \mathrm{Grad}(x_k) = \operatorname*{argmin}_{x \in Q}\{\langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2}\|x - x_k\|^2\}$$

If $Q = \mathbb{R}^n$ and ℓ_2-norm, $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$.

- **Primal progress**: Av. $\nabla f(x') \geq \frac{\nabla f(x)}{2}$ for $x' = \alpha x_k + (1 - \alpha) x_{k+1}$

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|\nabla f(x_k)\|_*^2$$

Gradient Descent: one dimensional intuition.

Convexity:
$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x)$.

Gradient Descent: one dimensional intuition.

Convexity:
$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x)$. $\implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$

Gradient Descent: one dimensional intuition.

Convexity:

$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x)$. $\implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$

Also: $f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*)$

Gradient Descent: one dimensional intuition.

Convexity:

$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$

Also: $f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*)$

L-Lipschitz, $R = \|x_0 - x^*\|$:

Gradient Descent: one dimensional intuition.

Convexity:

$f(x^*) \geq f(x) + \nabla f(x)^T(x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T(x - x^*)$

Also: $f(x) - f(x^*) \leq \nabla f(x)^T(x - x^*)$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$x^+ = x - \frac{1}{L}\nabla f(x)$

Gradient Descent: one dimensional intuition.

Convexity:

$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$

Also: $f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*)$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$x^+ = x - \frac{1}{L}\nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$

Gradient Descent: one dimensional intuition.

Convexity:

$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$

Also: $f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*)$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$x^+ = x - \frac{1}{L}\nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$

In one dimension: $\nabla f(x) = g$.

Gradient Descent: one dimensional intuition.

Convexity:

$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x)$. $\implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$

Also: $f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*)$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$x^+ = x - \frac{1}{L}\nabla f(x)$ $\qquad$ $f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$

In one dimension: $\nabla f(x) = g$.

Gap: gR.

Gradient Descent: one dimensional intuition.

Convexity:
$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$

Also: $f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*)$

L-Lipschitz, $R = \|x_0 - x^*\|$:
$x^+ = x - \frac{1}{L}\nabla f(x)$ $\quad$ $f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$

In one dimension: $\nabla f(x) = g$.

Gap: gR. Progress/step: Roughly $g^2/2$.

Gradient Descent: one dimensional intuition.

Convexity:

$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$

Also: $f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*)$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$x^+ = x - \frac{1}{L}\nabla f(x)$ $\qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$

In one dimension: $\nabla f(x) = g$.

Gap: gR. Progress/step: Roughly $g^2/2$.

Idea: Gap/(progress/step) $\implies$ roughly $2LR/g$ steps.

Gradient Descent: one dimensional intuition.

Convexity:

$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x).\ \implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$

Also: $f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*) = gR$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$x^+ = x - \frac{1}{L}\nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$

In one dimension: $\nabla f(x) = g$.

Gap: gR. Progress/step: Roughly $g^2/2$.

Idea: Gap/(progress/step) $\implies$ roughly $2LR/g$ steps.

Convexity: $g \geq (f(x) - f(x^*))/R$

Gradient Descent: one dimensional intuition.

Convexity:
$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$

Also: $f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*) = gR$

L-Lipschitz, $R = \|x_0 - x^*\|$:
$x^+ = x - \frac{1}{L}\nabla f(x)$ $\quad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$

In one dimension: $\nabla f(x) = g$.

Gap: gR. Progress/step: Roughly $g^2/2$.

Idea: Gap/(progress/step) $\implies$ roughly $2LR/g$ steps.

Convexity: $g \geq (f(x) - f(x^*))/R \implies 2LR^2/(f(x) - f(x^*))$ steps.

Gradient Descent: one dimensional intuition.

Convexity:
$f(x^*) \geq f(x) + \nabla f(x)^T(x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T(x - x^*)$

Also: $f(x) - f(x^*) \leq \nabla f(x)^T(x - x^*) = gR$

L-Lipschitz, $R = \|x_0 - x^*\|$:
$x^+ = x - \frac{1}{L}\nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$

In one dimension: $\nabla f(x) = g$.

Gap: gR. Progress/step: Roughly $g^2/2$.

Idea: Gap/(progress/step) $\implies$ roughly $2LR/g$ steps.

Convexity: $g \geq (f(x) - f(x^*))/R \implies 2LR^2/(f(x) - f(x^*))$ steps.

While gap $f(x) - f(x^*) \geq \varepsilon$ we have $g \geq \varepsilon/R$.

Gradient Descent: one dimensional intuition.

Convexity:
$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$

Also: $f(x) - f(x^*) \leq \nabla f(x)^T (x - x^*) = gR$

L-Lipschitz, $R = \|x_0 - x^*\|$:
$x^+ = x - \frac{1}{L}\nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$

In one dimension: $\nabla f(x) = g$.

Gap: gR. Progress/step: Roughly $g^2/2$.

Idea: Gap/(progress/step) $\implies$ roughly $2LR/g$ steps.

Convexity: $g \geq (f(x) - f(x^*))/R \implies 2LR^2/(f(x) - f(x^*))$ steps.

While gap $f(x) - f(x^*) \geq \varepsilon$ we have $g \geq \varepsilon/R$.

$\implies O(LR^2/\varepsilon)$ steps reduce gap by $1/2$.

Gradient Descent: convergence in ℓ_2

Convexity:
$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x)$.

Gradient Descent: convergence in ℓ_2

Convexity:

$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x)$. $\implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$

Gradient Descent: convergence in ℓ_2

Convexity:

$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x)$. $\implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$

L-Lipschitz, $R = \|x_0 - x^*\|$:

Gradient Descent: convergence in ℓ_2

Convexity:
$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$

L-Lipschitz, $R = \|x_0 - x^*\|$:
$x^+ = x - \frac{1}{L} \nabla f(x)$

Gradient Descent: convergence in ℓ_2

Convexity:

$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$x^+ = x - \frac{1}{L}\nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$

Gradient Descent: convergence in ℓ_2

Convexity:

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$$x^+ = x - \frac{1}{L}\nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$$

$$f(x^+) \leq f(x^*) + \nabla f(x)^T (x - x^*) - \frac{1}{2L}\|\nabla f(x)\|_2^2$$

Gradient Descent: convergence in ℓ_2

Convexity:

$f(x^*) \geq f(x) + \nabla f(x)^T(x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T(x - x^*)$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$x^+ = x - \frac{1}{L}\nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$

$f(x^+) \leq f(x^*) + \nabla f(x)^T(x - x^*) - \frac{1}{2L}\|\nabla f(x)\|_2^2$

Gradient Descent: convergence in ℓ_2

Convexity:

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$$x^+ = x - \frac{1}{L} \nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L} \|\nabla f(x)\|_*^2$$

$$f(x^+) \leq f(x^*) + \nabla f(x)^T (x - x^*) - \frac{1}{2L} \|\nabla f(x)\|_2^2$$

$$\implies f(x^+) - f(x^*) \leq \frac{L}{2} \left(\frac{2}{L} \nabla f(x)^T (x - x^*) - \frac{1}{L^2} \|\nabla f(x)\|_2^2 \right)$$

Gradient Descent: convergence in ℓ_2

Convexity:

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$$x^+ = x - \frac{1}{L}\nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$$

$$f(x^+) \leq f(x^*) + \nabla f(x)^T (x - x^*) - \frac{1}{2L}\|\nabla f(x)\|_2^2$$

$$\implies f(x^+) - f(x^*) \leq \frac{L}{2}\left(\frac{2}{L}\nabla f(x)^T (x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2\right)$$

$$\leq \frac{L}{2}\left(\frac{2}{L}\nabla f(x)^T (x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2 - \|x - x^*\|_2^2 + \|x - x^*\|_2^2\right) \text{ Add 0}$$

Gradient Descent: convergence in ℓ_2

Convexity:

$$f(x^*) \geq f(x) + \nabla f(x)^T(x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T(x - x^*)$$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$$x^+ = x - \frac{1}{L}\nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$$

$$f(x^+) \leq f(x^*) + \nabla f(x)^T(x - x^*) - \frac{1}{2L}\|\nabla f(x)\|_2^2$$

$$\implies f(x^+) - f(x^*) \leq \frac{L}{2}\left(\frac{2}{L}\nabla f(x)^T(x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2\right)$$

$$\leq \frac{L}{2}\left(\frac{2}{L}\nabla f(x)^T(x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2 - \|x - x^*\|_2^2 + \|x - x^*\|_2^2\right) \text{ Add 0}$$

$$\leq \frac{L}{2}(\|x - x^*\|_2^2 -$$

Gradient Descent: convergence in ℓ_2

Convexity:

$$f(x^*) \geq f(x) + \nabla f(x)^T(x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T(x - x^*)$$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$$x^+ = x - \frac{1}{L}\nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$$

$$f(x^+) \leq f(x^*) + \nabla f(x)^T(x - x^*) - \frac{1}{2L}\|\nabla f(x)\|_2^2$$

$$\implies f(x^+) - f(x^*) \leq \frac{L}{2}\left(\frac{2}{L}\nabla f(x)^T(x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2\right)$$

$$\leq \frac{L}{2}\left(\frac{2}{L}\nabla f(x)^T(x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2 - \|x - x^*\|_2^2 + \|x - x^*\|_2^2\right) \text{ Add 0}$$

$$\leq \frac{L}{2}\left(\|x - x^*\|_2^2 - \|(x - x^*) - \frac{1}{L}\nabla f(x)\|_2^2\right)$$

$$\leq \frac{L}{2}\left(\|x - x^*\|_2^2 - \right.$$

Gradient Descent: convergence in ℓ_2

Convexity:

$$f(x^*) \geq f(x) + \nabla f(x)^T(x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T(x - x^*)$$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$$x^+ = x - \frac{1}{L}\nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$$

$$f(x^+) \leq f(x^*) + \nabla f(x)^T(x - x^*) - \frac{1}{2L}\|\nabla f(x)\|_2^2$$

$$\implies f(x^+) - f(x^*) \leq \frac{L}{2}\left(\frac{2}{L}\nabla f(x)^T(x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2\right)$$

$$\leq \frac{L}{2}\left(\frac{2}{L}\nabla f(x)^T(x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2 - \|x - x^*\|_2^2 + \|x - x^*\|_2^2\right) \text{ Add 0}$$

$$\leq \frac{L}{2}\left(\|x - x^*\|_2^2 - \|(x - x^*) - \frac{1}{L}\nabla f(x)\|_2^2\right)$$

$$\leq \frac{L}{2}\left(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2\right)$$

Gradient Descent: convergence in ℓ_2

Convexity:
$$f(x^*) \geq f(x) + \nabla f(x)^T(x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T(x - x^*)$$

L-Lipschitz, $R = \|x_0 - x^*\|$:
$$x^+ = x - \frac{1}{L}\nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$$

$$f(x^+) \leq f(x^*) + \nabla f(x)^T(x - x^*) - \frac{1}{2L}\|\nabla f(x)\|_2^2$$

$$\implies f(x^+) - f(x^*) \leq \frac{L}{2}(\frac{2}{L}\nabla f(x)^T(x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2)$$

$$\leq \frac{L}{2}(\frac{2}{L}\nabla f(x)^T(x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2 - \|x - x^*\|_2^2 + \|x - x^*\|_2^2) \text{ Add 0}$$

$$\leq \frac{L}{2}(\|x - x^*\|_2^2 - \|(x - x^*) - \frac{1}{L}\nabla f(x)\|_2^2)$$
$$\leq \frac{L}{2}(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2)$$

$$\sum_k^T f(x_k) - f(x^*) \leq \sum_k^T \frac{L}{2}(\|x_{k-1} - x^*\|_2^2 - \|x_k - x^*\|_2^2)$$

Gradient Descent: convergence in ℓ_2

Convexity:

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$$x^+ = x - \frac{1}{L}\nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$$

$$f(x^+) \leq f(x^*) + \nabla f(x)^T (x - x^*) - \frac{1}{2L}\|\nabla f(x)\|_2^2$$

$$\implies f(x^+) - f(x^*) \leq \frac{L}{2}\left(\frac{2}{L}\nabla f(x)^T (x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2\right)$$

$$\leq \frac{L}{2}\left(\frac{2}{L}\nabla f(x)^T (x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2 - \|x - x^*\|_2^2 + \|x - x^*\|_2^2\right) \text{ Add 0}$$

$$\leq \frac{L}{2}\left(\|x - x^*\|_2^2 - \|(x - x^*) - \frac{1}{L}\nabla f(x)\|_2^2\right)$$

$$\leq \frac{L}{2}\left(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2\right)$$

$$\sum_k^T f(x_k) - f(x^*) \leq \sum_k^T \frac{L}{2}\left(\|x_{k-1} - x^*\|_2^2 - \|x_k - x^*\|_2^2\right)$$

$$\leq \frac{L}{2}\left(\|x_0 - x^*\|_2^2 - \|x_T - x^*\|_2^2\right) \leq \frac{L}{2}\|x_0 - x^*\|_2^2$$

Gradient Descent: convergence in ℓ_2

Convexity:

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$$x^+ = x - \frac{1}{L}\nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$$

$$f(x^+) \leq f(x^*) + \nabla f(x)^T (x - x^*) - \frac{1}{2L}\|\nabla f(x)\|_2^2$$

$$\implies f(x^+) - f(x^*) \leq \frac{L}{2}\left(\frac{2}{L}\nabla f(x)^T (x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2\right)$$

$$\leq \frac{L}{2}\left(\frac{2}{L}\nabla f(x)^T (x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2 - \|x - x^*\|_2^2 + \|x - x^*\|_2^2\right) \text{ Add 0}$$

$$\leq \frac{L}{2}\left(\|x - x^*\|_2^2 - \|(x - x^*) - \frac{1}{L}\nabla f(x)\|_2^2\right)$$
$$\leq \frac{L}{2}\left(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2\right)$$

$$\sum_k^T f(x_k) - f(x^*) \leq \sum_k^T \frac{L}{2}\left(\|x_{k-1} - x^*\|_2^2 - \|x_k - x^*\|_2^2\right)$$
$$\leq \frac{L}{2}\left(\|x_0 - x^*\|_2^2 - \|x_T - x^*\|_2^2\right) \leq \frac{L}{2}\|x_0 - x^*\|_2^2$$

Gradient Descent: convergence in ℓ_2

Convexity:

$$f(x^*) \geq f(x) + \nabla f(x)^T(x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T(x - x^*)$$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$$x^+ = x - \frac{1}{L}\nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$$

$$f(x^+) \leq f(x^*) + \nabla f(x)^T(x - x^*) - \frac{1}{2L}\|\nabla f(x)\|_2^2$$

$$\implies f(x^+) - f(x^*) \leq \frac{L}{2}\left(\frac{2}{L}\nabla f(x)^T(x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2\right)$$

$$\leq \frac{L}{2}\left(\frac{2}{L}\nabla f(x)^T(x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2 - \|x - x^*\|_2^2 + \|x - x^*\|_2^2\right) \text{ Add 0}$$

$$\leq \frac{L}{2}\left(\|x - x^*\|_2^2 - \|(x - x^*) - \frac{1}{L}\nabla f(x)\|_2^2\right)$$
$$\leq \frac{L}{2}\left(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2\right)$$

$$\sum_k^T f(x_k) - f(x^*) \leq \sum_k^T \frac{L}{2}\left(\|x_{k-1} - x^*\|_2^2 - \|x_k - x^*\|_2^2\right)$$
$$\leq \frac{L}{2}\left(\|x_0 - x^*\|_2^2 - \|x_T - x^*\|_2^2\right) \leq \frac{L}{2}\|x_0 - x^*\|_2^2$$

$f(x_k)$ is decreasing, we have $f(x_T) \leq \frac{1}{T}\sum_k f(x_k)$.

Gradient Descent: convergence in ℓ_2

Convexity:

$$f(x^*) \geq f(x) + \nabla f(x)^T(x^* - x). \implies f(x) \leq f(x^*) + \nabla f(x)^T(x - x^*)$$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$$x^+ = x - \frac{1}{L}\nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$$

$$f(x^+) \leq f(x^*) + \nabla f(x)^T(x - x^*) - \frac{1}{2L}\|\nabla f(x)\|_2^2$$

$$\implies f(x^+) - f(x^*) \leq \frac{L}{2}\left(\frac{2}{L}\nabla f(x)^T(x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2\right)$$

$$\leq \frac{L}{2}\left(\frac{2}{L}\nabla f(x)^T(x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2 - \|x - x^*\|_2^2 + \|x - x^*\|_2^2\right) \text{ Add 0}$$

$$\leq \frac{L}{2}\left(\|x - x^*\|_2^2 - \|(x - x^*) - \frac{1}{L}\nabla f(x)\|_2^2\right)$$
$$\leq \frac{L}{2}\left(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2\right)$$

$$\sum_k^T f(x_k) - f(x^*) \leq \sum_k^T \frac{L}{2}\left(\|x_{k-1} - x^*\|_2^2 - \|x_k - x^*\|_2^2\right)$$
$$\leq \frac{L}{2}\left(\|x_0 - x^*\|_2^2 - \|x_T - x^*\|_2^2\right) \leq \frac{L}{2}\|x_0 - x^*\|_2^2$$

$f(x_k)$ is decreasing, we have $f(x_T) \leq \frac{1}{T}\sum_k f(x_k)$.

$\implies f(x_T) - f(x^*) \leq \frac{LR^2}{2T}$ where $R = \|x_0 - x^*\|$.

Gradient Descent: convergence in ℓ_2

Convexity:

$f(x^*) \geq f(x) + \nabla f(x)^T(x^* - x)$. $\implies f(x) \leq f(x^*) + \nabla f(x)^T(x - x^*)$

L-Lipschitz, $R = \|x_0 - x^*\|$:

$x^+ = x - \frac{1}{L}\nabla f(x) \qquad f(x) - f(x^+) \geq \frac{1}{2L}\|\nabla f(x)\|_*^2$

$f(x^+) \leq f(x^*) + \nabla f(x)^T(x - x^*) - \frac{1}{2L}\|\nabla f(x)\|_2^2$

$\implies f(x^+) - f(x^*) \leq \frac{L}{2}(\frac{2}{L}\nabla f(x)^T(x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2)$

$\leq \frac{L}{2}(\frac{2}{L}\nabla f(x)^T(x - x^*) - \frac{1}{L^2}\|\nabla f(x)\|_2^2 - \|x - x^*\|_2^2 + \|x - x^*\|_2^2)$ Add 0

$\leq \frac{L}{2}(\|x - x^*\|_2^2 - \|(x - x^*) - \frac{1}{L}\nabla f(x)\|_2^2)$

$\leq \frac{L}{2}(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2)$

$\sum_k^T f(x_k) - f(x^*) \leq \sum_k^T \frac{L}{2}(\|x_{k-1} - x^*\|_2^2 - \|x_k - x^*\|_2^2)$

$\leq \frac{L}{2}(\|x_0 - x^*\|_2^2 - \|x_T - x^*\|_2^2) \leq \frac{L}{2}\|x_0 - x^*\|_2^2$

$f(x_k)$ is decreasing, we have $f(x_T) \leq \frac{1}{T}\sum_k f(x_k)$.

$\implies f(x_T) - f(x^*) \leq \frac{LR^2}{2T}$ where $R = \|x_0 - x^*\|$.

Also: $T = O(LR^2/\varepsilon)$ iterations for $f(x_T) - f(x^*) \leq \varepsilon$.

Gradient Descent

Primal progress

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L} \|\nabla f(x)\|_*^2$$

Convergence

L-Lipschitz, $R = \max_{x: f(x) \leq f(x_0)} \|x - x^*\|$:

$$f(x_T) - f(x^*) \leq O(\frac{LR^2}{T})$$

To get ε-approximation, need

$$T = O(\frac{LR^2}{\varepsilon})$$

Relationship?

What is relationship to move closer to feasible?

Relationship?

What is relationship to move closer to feasible?

If wrong side of hyperplane by at least something.

Relationship?

What is relationship to move closer to feasible?

If wrong side of hyperplane by at least something.
Move to other side.

Relationship?

What is relationship to move closer to feasible?

If wrong side of hyperplane by at least something.
Move to other side.

Relationship?

What is relationship to move closer to feasible?

If wrong side of hyperplane by at least something.
Move to other side.

What is the "hyperplane" here?

Relationship?

What is relationship to move closer to feasible?

If wrong side of hyperplane by at least something.
Move to other side.

What is the "hyperplane" here?

$\nabla f(x)$

Relationship?

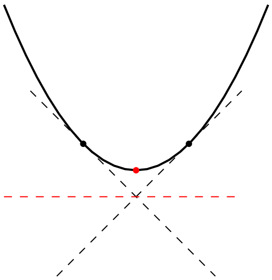
What is relationship to move closer to feasible?

If wrong side of hyperplane by at least something.
Move to other side.

What is the "hyperplane" here?

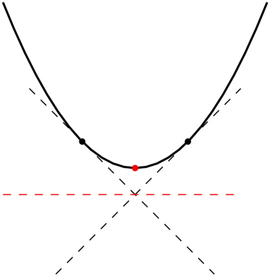
$\nabla f(x)$ Maybe.

Mirror Descent



- Each point gives a linear lower bound.
- Average of the lower bounds becomes flatter.
- Add the point with current worst regret.
- Output average of queried points.
- $x = \alpha x_1 + (1 - \alpha) x_2$

Mirror Descent



- Each point gives a linear lower bound.
- Average of the lower bounds becomes flatter.
- Add the point with current worst regret.
- Output average of queried points.
- $x = \alpha x_1 + (1-\alpha)x_2$
  $\implies f(x) \le \alpha f(x_1) + (1-\alpha)f(x_2)$.

Analysis doesn't require L-Lipschitz.

Mirror Descent: Regret Minimization

- Average **Regret** with loss vector ξ_i's

$$R_k(u) = \frac{1}{k}\sum_{i=0}^{k-1}\langle \xi_i, z_i - u\rangle$$

Why care about average regret? Bounds gap to OPT:

With $\xi_i = \nabla f(z_i)$, $\bar{z} = \frac{1}{k}\sum_{i=0}^{k-1} z_i$,

$$f(\bar{z}) - f(u) \le \frac{1}{k}\sum_{i=0}^{k-1} f(z_i) - f(u) \le \frac{1}{k}\sum_{i=0}^{k-1}\langle \nabla f(z_i), z_i - u\rangle = R_k(u)$$

$$f(\bar{z}) - \mathrm{OPT} \le \max_u R_k(u)$$

Mirror Descent: Regret Minimization

- Regularized average regret

$$\tilde{R}_k(u) = \frac{1}{\alpha k}(-w(u) + \alpha \sum_{i=0}^{k-1} \langle \xi_i, z_i - u \rangle)$$
$$= R_k(u) - \frac{w(u)}{\alpha k}$$

Distance Generating Function

$$w : \mathbb{R}^n \to \mathbb{R}$$

1-strongly convex for norm $\|\cdot\|$:

$$w(y) \geq w(x) + \langle \nabla w(x), y - x \rangle + \frac{1}{2}\|x - y\|^2$$

Mirror Descent: Regret Minimization

- Regularized average regret

$$\tilde{R}_k(u) = \frac{1}{\alpha k}(-w(u) + \alpha \sum_{i=0}^{k-1} \langle \xi_i, z_i - u \rangle)$$
$$= R_k(u) - \frac{w(u)}{\alpha k}$$

Distance Generating Function

$$w : \mathbb{R}^n \to \mathbb{R}$$

1-strongly convex for norm $\|\cdot\|$:

$$w(y) \geq w(x) + \langle \nabla w(x), y - x \rangle + \frac{1}{2}\|x - y\|^2$$

For ℓ_2-norm, simply $w(x) = \frac{1}{2}\|x\|_2^2$.

Mirror Descent: Regret Minimization

- Regularized average regret

$$\tilde{R}_k(u) = \frac{1}{\alpha k}\left(-w(u) + \alpha \sum_{i=0}^{k-1} \langle \xi_i, z_i - u \rangle\right)$$
$$= R_k(u) - \frac{w(u)}{\alpha k}$$

Distance Generating Function

$$w : \mathbb{R}^n \to \mathbb{R}$$

1-strongly convex for norm $\|\cdot\|$:

$$w(y) \geq w(x) + \langle \nabla w(x), y - x \rangle + \frac{1}{2}\|x - y\|^2$$

For ℓ_2-norm, simply $w(x) = \frac{1}{2}\|x\|_2^2$.
(For distributions: $w(x) = -\sum_i x_i \log x_i$.)

Bregman divergence

$$V_x(y) = w(y) - \langle \nabla w(x), y - x \rangle - w(x) \geq \frac{1}{2}\|x - y\|^2$$

Standard three point property of Bregman divergence:

$$\forall x, y \geq 0 \quad \langle -\nabla V_x(y), y - u \rangle = V_x(u) - V_y(u) - V_x(y),$$

Bregman divergence

$$V_x(y) = w(y) - \langle \nabla w(x), y - x \rangle - w(x) \geq \frac{1}{2}\|x - y\|^2$$

Standard three point property of Bregman divergence:

$$\forall x, y \geq 0 \quad \langle -\nabla V_x(y), y - u \rangle = V_x(u) - V_y(u) - V_x(y),$$

For ℓ_2-norm, $V_x(y) = \frac{1}{2}\|x - y\|_2^2$, $\nabla V_x(y) = (x - y)$

Bregman divergence

$$V_x(y) = w(y) - \langle \nabla w(x), y - x \rangle - w(x) \geq \frac{1}{2}\|x - y\|^2$$

Standard three point property of Bregman divergence:

$$\forall x, y \geq 0 \quad \langle -\nabla V_x(y), y - u \rangle = V_x(u) - V_y(u) - V_x(y),$$

For ℓ_2-norm, $V_x(y) = \frac{1}{2}\|x - y\|_2^2$, $\nabla V_x(y) = (x - y)$

Three point property $\leftrightarrow$ Law of cosines

Bregman divergence

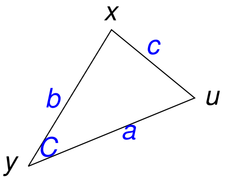
$$V_x(y) = w(y) - \langle \nabla w(x), y - x \rangle - w(x) \geq \frac{1}{2}\|x - y\|^2$$

Standard three point property of Bregman divergence:

$$\forall x, y \geq 0 \quad \langle -\nabla V_x(y), y - u \rangle = V_x(u) - V_y(u) - V_x(y),$$

For ℓ_2-norm, $V_x(y) = \frac{1}{2}\|x - y\|_2^2$, $\nabla V_x(y) = (x - y)$

Three point property $\leftrightarrow$ Law of cosines



$$c^2 = a^2 + b^2 - 2ab\cos(C)$$

Bregman divergence

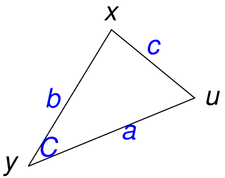
$$V_x(y) = w(y) - \langle \nabla w(x), y - x \rangle - w(x) \geq \frac{1}{2}\|x - y\|^2$$

Standard three point property of Bregman divergence:

$$\forall x, y \geq 0 \quad \langle -\nabla V_x(y), y - u \rangle = V_x(u) - V_y(u) - V_x(y),$$

For ℓ_2-norm, $V_x(y) = \frac{1}{2}\|x - y\|_2^2$, $\nabla V_x(y) = (x - y)$

Three point property $\leftrightarrow$ Law of cosines



$$c^2 = a^2 + b^2 - 2ab\cos(C) \text{ or } 2ab\cos(C) = c^2 - a^2 - b^2$$

Bregman divergence

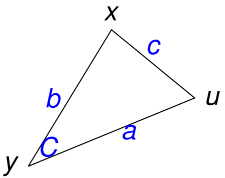
$$V_x(y) = w(y) - \langle \nabla w(x), y - x \rangle - w(x) \geq \frac{1}{2}\|x - y\|^2$$

Standard three point property of Bregman divergence:

$$\forall x, y \geq 0 \quad \langle -\nabla V_x(y), y - u \rangle = V_x(u) - V_y(u) - V_x(y),$$

For ℓ_2-norm, $V_x(y) = \frac{1}{2}\|x - y\|_2^2$, $\nabla V_x(y) = (x - y)$

Three point property $\leftrightarrow$ Law of cosines



$$c^2 = a^2 + b^2 - 2ab\cos(C) \text{ or } 2ab\cos(C) = c^2 - a^2 - b^2$$

$$a^2 = V_y(u), \; b^2 = V_x(u), \quad c^2 = V_x(u), \; 2ab\cos(C) = -(x - y) \cdot (y - u)$$

Mirror Descent

$$z_{k+1} = \mathrm{Mirr}(z_k, \alpha\xi_k) = \operatorname*{argmin}_{z \in Q}\{V_{z_k}(z) + \alpha\langle \xi_k, z - z_k\rangle\}$$

Mirror Descent

$$z_{k+1} = \mathrm{Mirr}(z_k, \alpha\xi_k) = \operatorname*{argmin}_{z \in Q} \{ V_{z_k}(z) + \alpha \langle \xi_k, z - z_k \rangle \}$$

Equivalent to regret minimization when $Q = \mathbb{R}^n$:

- Optimality condition of MD step:

$$\begin{aligned} \nabla V_{z_k}(z_{k+1}) &= -\alpha\xi_k \\ z_{k+1} - z_k &= -\alpha\xi_k \\ z_{k+1} &= z_0 - \sum_i \alpha\xi_i \end{aligned}$$

- Regret Minimization:

$$z_{k+1} = \operatorname*{argmax}_{z} \{ -w(z) + \alpha \sum_{i=0}^{k} \langle \xi_i, z_i - z \rangle \}$$

Optimality condition:

$$z_{k+1} = -\sum_i \alpha\xi_i$$

Mirror Descent

$$z_{k+1} = \mathrm{Mirr}(z_k, \alpha\xi_k) = \operatorname*{argmin}_{z \in Q}\{V_{z_k}(z) + \alpha\langle \xi_k, z - z_k\rangle\}$$

Recall: $\nabla V_{z_k}(z_{k+1}) = -\alpha\xi_k \quad z_{k+1} - z_k = -\alpha\xi_k$

Lemma:

$$\begin{aligned}\alpha\langle \xi_k, z_k - u\rangle &\leq \alpha\langle \xi_k, z_k - z_{k+1}\rangle + V_{z_k}(u) - V_{z_{k+1}}(u) - V_{z_k}(z_{k+1}) \\ &\leq \frac{\alpha^2}{2}\|\xi_k\|_*^2 + V_{z_k}(u) - V_{z_{k+1}}(u) \qquad \forall u \in Q\end{aligned}$$

$$\begin{aligned}&= \alpha\langle \xi_k, z_k - z_{k+1}\rangle - \langle \nabla V_{z_k}(z_{k+1}), z_{k+1} - u\rangle \\ &= \alpha\langle \xi_k, z_k - z_{k+1}\rangle + V_{z_k}(u) - V_{z_{k+1}}(u) - V_{z_k}(z_{k+1})\end{aligned}$$

Mirror Descent

$$z_{k+1} = \mathrm{Mirr}(z_k, \alpha\xi_k) = \operatorname*{argmin}_{z \in Q}\{V_{z_k}(z) + \alpha\langle \xi_k, z - z_k\rangle\}$$

Recall: $\nabla V_{z_k}(z_{k+1}) = -\alpha\xi_k \quad z_{k+1} - z_k = -\alpha\xi_k$

Lemma:

$$\begin{aligned}\alpha\langle \xi_k, z_k - u\rangle &\leq \alpha\langle \xi_k, z_k - z_{k+1}\rangle + V_{z_k}(u) - V_{z_{k+1}}(u) - V_{z_k}(z_{k+1}) \\ &\leq \frac{\alpha^2}{2}\|\xi_k\|_*^2 + V_{z_k}(u) - V_{z_{k+1}}(u) \qquad \forall u \in Q\end{aligned}$$

Proof:

$$\alpha\langle \xi_k, z_k - u\rangle = \alpha\langle \xi_k, z_k - z_{k+1}\rangle + \alpha\langle \xi_k, z_{k+1} - u\rangle$$

□

Mirror Descent

$$z_{k+1} = \mathrm{Mirr}(z_k, \alpha\xi_k) = \operatorname*{argmin}_{z\in Q}\{V_{z_k}(z) + \alpha\langle \xi_k, z - z_k\rangle\}$$

Recall: $\nabla V_{z_k}(z_{k+1}) = -\alpha\xi_k \quad z_{k+1} - z_k = -\alpha\xi_k$

Lemma:

$$\begin{aligned}\alpha\langle \xi_k, z_k - u\rangle &\le \alpha\langle \xi_k, z_k - z_{k+1}\rangle + V_{z_k}(u) - V_{z_{k+1}}(u) - V_{z_k}(z_{k+1})\\ &\le \frac{\alpha^2}{2}\|\xi_k\|_*^2 + V_{z_k}(u) - V_{z_{k+1}}(u) \qquad \forall u \in Q\end{aligned}$$

Proof:

$$\begin{aligned}\alpha\langle \xi_k, z_k - u\rangle &= \alpha\langle \xi_k, z_k - z_{k+1}\rangle + \alpha\langle \xi_k, z_{k+1} - u\rangle\\ &= \alpha\langle \xi_k, z_k - z_{k+1}\rangle - \langle \nabla V_{z_k}(z_{k+1}), z_{k+1} - u\rangle\\ &= \alpha\langle \xi_k, z_k - z_{k+1}\rangle + V_{z_k}(u) - V_{z_{k+1}}(u) - V_{z_k}(z_{k+1})\\ &\le \frac{\alpha^2}{2}\|\xi_k\|_*^2 + V_{z_k}(u) - V_{z_{k+1}}(u)\end{aligned}$$

□

Mirror Descent

- $\alpha\langle \xi_k, z_k - u\rangle \leq \frac{\alpha^2}{2}\|\xi_k\|_*^2 + V_{z_k}(u) - V_{z_{k+1}}(u)$

  Telescoping T iterations, and **width** $\|\xi_k\|_*^2 \leq \rho^2$

Mirror Descent

- $\alpha\langle\xi_k, z_k - u\rangle \leq \frac{\alpha^2}{2}\|\xi_k\|_*^2 + V_{z_k}(u) - V_{z_{k+1}}(u)$

  Telescoping T iterations, and **width** $\|\xi_k\|_*^2 \leq \rho^2$

$$1 \cdot \frac{\alpha^2\rho^2}{2} + V_{z_0}(u) - V_{z_1}(u)$$

Mirror Descent

- $\alpha\langle \xi_k, z_k - u\rangle \leq \frac{\alpha^2}{2}\|\xi_k\|_*^2 + V_{z_k}(u) - V_{z_{k+1}}(u)$

  Telescoping T iterations, and **width** $\|\xi_k\|_*^2 \leq \rho^2$

$$2 \cdot \frac{\alpha^2\rho^2}{2} + V_{z_0}(u) - \cancel{V_{z_1}(u)} + \cancel{V_{z_1}(u)} - V_{z_2}(u)$$

Mirror Descent

- $\alpha\langle \xi_k, z_k - u\rangle \le \frac{\alpha^2}{2}\|\xi_k\|_*^2 + V_{z_k}(u) - V_{z_{k+1}}(u)$

  Telescoping T iterations, and **width** $\|\xi_k\|_*^2 \le \rho^2$

  $$3 \cdot \frac{\alpha^2\rho^2}{2} + V_{z_0}(u) - \cancel{V_{z_1}(u)} + \cancel{V_{z_1}(u)} - \cancel{V_{z_2}(u)} + \cancel{V_{z_2}(u)} - V_{z_3}(u) \ldots$$

Mirror Descent

- $\alpha\langle \xi_k, z_k - u\rangle \le \frac{\alpha^2}{2}\|\xi_k\|_*^2 + V_{z_k}(u) - V_{z_{k+1}}(u)$

  Telescoping T iterations, and **width** $\|\xi_k\|_*^2 \le \rho^2$

$$\alpha \sum_{i=0}^{T-1} \langle \xi_i, z_i - u\rangle \le \frac{\alpha^2 \rho^2 T}{2} + V_{z_0}(u) - V_{z_T}(u)$$

Mirror Descent

- $\alpha\langle \xi_k, z_k - u\rangle \le \frac{\alpha^2}{2}\|\xi_k\|_*^2 + V_{z_k}(u) - V_{z_{k+1}}(u)$

  Telescoping T iterations, and **width** $\|\xi_k\|_*^2 \le \rho^2$

  $$\alpha \sum_{i=0}^{T-1} \langle \xi_i, z_i - u\rangle \le \frac{\alpha^2 \rho^2 T}{2} + V_{z_0}(u) - V_{z_T}(u)$$

- $\alpha = \frac{\varepsilon}{\rho^2}$, **diameter** $V_{z_0}(u) \le \Theta$, in $T = \frac{2\rho^2\Theta}{\varepsilon^2}$ iterations

  $$\forall u, f(\bar{z}) - f(u) \le \frac{\alpha\rho^2}{2} + \frac{V_{z_0}(u)}{\alpha T} \le \varepsilon$$

Mirror Descent

- $\alpha\langle \xi_k, z_k - u\rangle \leq \frac{\alpha^2}{2}\|\xi_k\|_*^2 + V_{z_k}(u) - V_{z_{k+1}}(u)$

  Telescoping T iterations, and **width** $\|\xi_k\|_*^2 \leq \rho^2$

$$\alpha \sum_{i=0}^{T-1} \langle \xi_i, z_i - u\rangle \leq \frac{\alpha^2 \rho^2 T}{2} + V_{z_0}(u) - V_{z_T}(u)$$

- $\alpha = \frac{\varepsilon}{\rho^2}$, **diameter** $V_{z_0}(u) \leq \Theta$, in $T = \frac{2\rho^2\Theta}{\varepsilon^2}$ iterations

$$\forall u, f(\bar{z}) - f(u) \leq \frac{\alpha\rho^2}{2} + \frac{V_{z_0}(u)}{\alpha T} \leq \varepsilon$$

- Regret terms $\frac{\alpha^2}{2}\|\xi_k\|_*^2$ accumulate, bound step size α.
  Bregman divergence terms telescope.

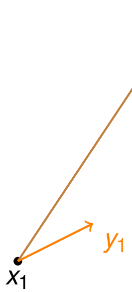
Linear Coupling

Intuition: If $\|\nabla f(x_k)\|_*^2$ large

- GD can make large primal progress $\frac{1}{2L}\|\nabla f(x_k)\|_*^2$
- MD suffers large regret $\frac{\alpha^2}{2}\|\nabla f(x)\|_*^2$
- Use primal progress to cover regret.
- Regret terms no longer accumulates,
  telescope as the primal progress.

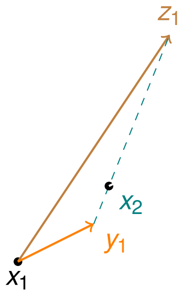
Linear Coupling

- $x_0 = y_0 = z_0$.
- Coupling: $x_{k+1} = \tau z_k + (1-\tau) y_k$.
- MD: $z_{k+1} = \mathrm{Mirr}(z_k, \alpha \nabla f(x_{k+1}))$
- GD: $y_{k+1} = \mathrm{Grad}(x_{k+1})$.



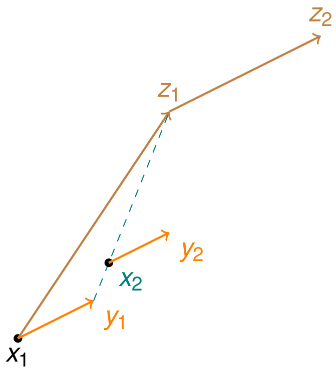
Linear Coupling

- $x_0 = y_0 = z_0$.
- Coupling: $x_{k+1} = \tau z_k + (1-\tau) y_k$.
- MD: $z_{k+1} = \mathrm{Mirr}(z_k, \alpha \nabla f(x_{k+1}))$
- GD: $y_{k+1} = \mathrm{Grad}(x_{k+1})$.



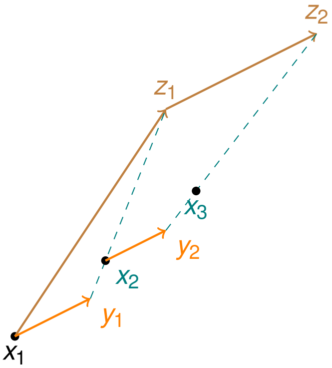
Linear Coupling

- $x_0 = y_0 = z_0$.
- Coupling: $x_{k+1} = \tau z_k + (1-\tau) y_k$.
- MD: $z_{k+1} = \mathrm{Mirr}(z_k, \alpha \nabla f(x_{k+1}))$
- GD: $y_{k+1} = \mathrm{Grad}(x_{k+1})$.



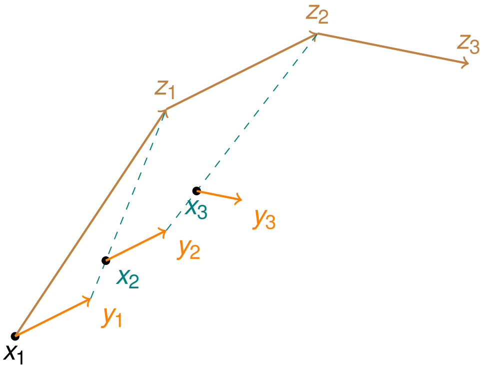
Linear Coupling

- $x_0 = y_0 = z_0$.
- Coupling: $x_{k+1} = \tau z_k + (1 - \tau) y_k$.
- MD: $z_{k+1} = \mathrm{Mirr}(z_k, \alpha \nabla f(x_{k+1}))$
- GD: $y_{k+1} = \mathrm{Grad}(x_{k+1})$.



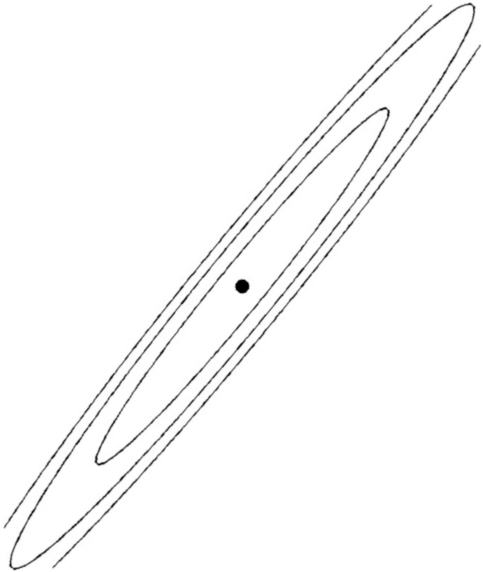
Linear Coupling

- $x_0 = y_0 = z_0$.
- Coupling: $x_{k+1} = \tau z_k + (1-\tau) y_k$.
- MD: $z_{k+1} = \mathrm{Mirr}(z_k, \alpha \nabla f(x_{k+1}))$
- GD: $y_{k+1} = \mathrm{Grad}(x_{k+1})$.



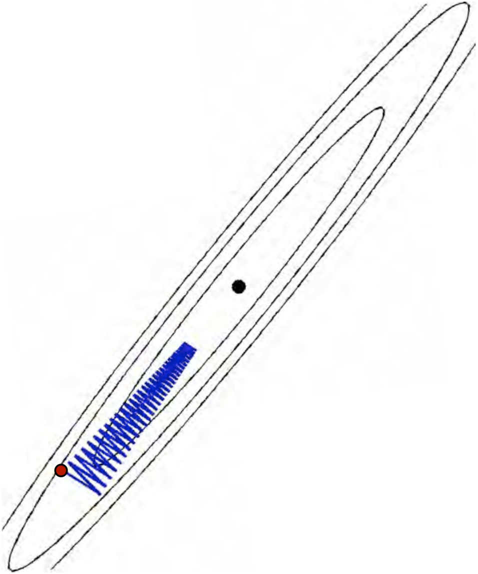
Linear Coupling

Momentum View:



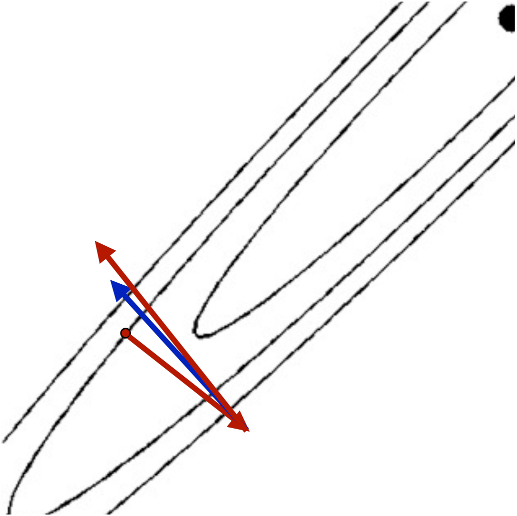
Linear Coupling

Momentum View:



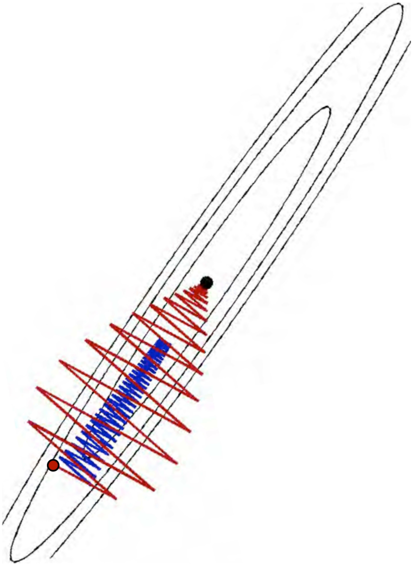
Linear Coupling

Momentum View:

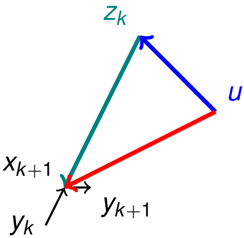


Linear Coupling

Momentum View:

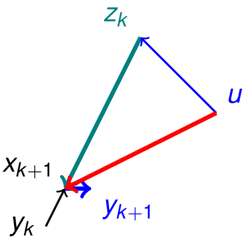


Bound $\alpha(f(x_{k+1}) - f(u)) \leq \alpha\langle\nabla f(x_{k+1}), x_{k+1} - u\rangle$



$$\begin{aligned} &\alpha\langle\nabla f(x_{k+1}), x_{k+1} - u\rangle \\ =&\alpha\langle\nabla f(x_{k+1}), z_k - u\rangle \\ &+ \alpha\langle\nabla f(x_{k+1}), x_{k+1} - z_k\rangle \end{aligned}$$

Bound $\alpha(f(x_{k+1}) - f(u)) \le \alpha\langle \nabla f(x_{k+1}), x_{k+1} - u\rangle$

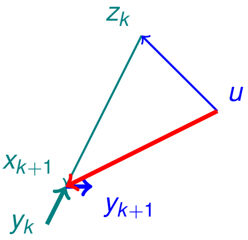


$$\begin{aligned} &\alpha\langle \nabla f(x_{k+1}), x_{k+1} - u\rangle \\ =&\alpha\langle \nabla f(x_{k+1}), z_k - u\rangle \\ &+ \alpha\langle \nabla f(x_{k+1}), x_{k+1} - z_k\rangle \\ \le&\alpha^2 L(f(x_{k+1}) - f(y_{k+1})) \\ &+ V_{z_k}(u) - V_{z_{k+1}}(u) \end{aligned}$$

MD: $\alpha\langle \nabla f(x_{k+1}), z_k - u\rangle \le \frac{\alpha^2}{2}\|\nabla f(x_{k+1})\|_*^2 + V_{z_k}(u) - V_{z_{k+1}}(u)$

GD: $f(x_{k+1}) - f(y_{k+1}) \ge \frac{1}{2L}\|\nabla f(x_{k+1})\|_*^2$

Bound $\alpha(f(x_{k+1}) - f(u)) \leq \alpha\langle \nabla f(x_{k+1}), x_{k+1} - u\rangle$

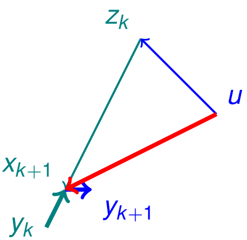


$$
\begin{aligned}
&\alpha\langle \nabla f(x_{k+1}), x_{k+1} - u\rangle \\
=&\alpha\langle \nabla f(x_{k+1}), z_k - u\rangle \\
&+ \alpha\langle \nabla f(x_{k+1}), x_{k+1} - z_k\rangle \\
\leq&\alpha^2 L(f(x_{k+1}) - f(y_{k+1})) \\
&+ V_{z_k}(u) - V_{z_{k+1}}(u) \\
&+ \alpha\langle \nabla f(x_{k+1}), \frac{1-\tau}{\tau}(y_k - x_{k+1})\rangle
\end{aligned}
$$

Coupling:

$x_{k+1} = \tau z_k + (1-\tau) y_k \quad \rightarrow \quad \tau(x_{k+1} - z_k) = (1-\tau)(y_k - x_{k+1})$

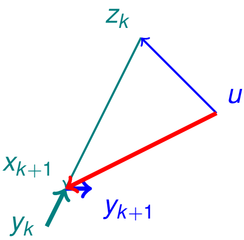
Bound $\alpha(f(x_{k+1}) - f(u)) \le \alpha\langle \nabla f(x_{k+1}), x_{k+1} - u\rangle$



$$
\begin{aligned}
&\alpha\langle \nabla f(x_{k+1}), x_{k+1} - u\rangle \\
=&\alpha\langle \nabla f(x_{k+1}), z_k - u\rangle \\
&+ \alpha\langle \nabla f(x_{k+1}), x_{k+1} - z_k\rangle \\
\le&\alpha^2 L(f(x_{k+1}) - f(y_{k+1})) \\
&+ V_{z_k}(u) - V_{z_{k+1}}(u) \\
&+ \frac{1-\tau}{\tau}\alpha(f(y_k) - f(x_{k+1}))
\end{aligned}
$$

$$
\begin{aligned}
\alpha\langle \nabla f(x_{k+1}), \tfrac{1-\tau}{\tau}(y_k - x_{k+1})\rangle &= \alpha\tfrac{1-\tau}{\tau}\langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle) \\
&\le \tfrac{1-\tau}{\tau}\alpha(f(y_k) - f(x_{k+1}))
\end{aligned}
$$

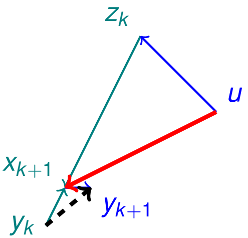
Bound $\alpha(f(x_{k+1}) - f(u)) \leq \alpha\langle \nabla f(x_{k+1}), x_{k+1} - u\rangle$



$$
\begin{aligned}
& \alpha\langle \nabla f(x_{k+1}), x_{k+1} - u\rangle \\
= & \alpha\langle \nabla f(x_{k+1}), z_k - u\rangle \\
& + \alpha\langle \nabla f(x_{k+1}), x_{k+1} - z_k\rangle \\
\leq & \alpha^2 L(f(x_{k+1}) - f(y_{k+1})) \\
& + V_{z_k}(u) - V_{z_{k+1}}(u) \\
& + \alpha^2 L(f(y_k) - f(x_{k+1}))
\end{aligned}
$$

Let $\alpha^2 L = \dfrac{1-\tau}{\tau}\alpha$

Bound $\alpha(f(x_{k+1}) - f(u)) \leq \alpha\langle\nabla f(x_{k+1}), x_{k+1} - u\rangle$



$$
\begin{aligned}
&\alpha\langle\nabla f(x_{k+1}), x_{k+1} - u\rangle \\
=&\alpha\langle\nabla f(x_{k+1}), z_k - u\rangle \\
&+ \alpha\langle\nabla f(x_{k+1}), x_{k+1} - z_k\rangle \\
\leq&\alpha^2 L(\cancel{f(x_{k+1})} - f(y_{k+1})) \\
&+ V_{z_k}(u) - V_{z_{k+1}}(u) \\
&+ \alpha^2 L(f(y_k) - \cancel{f(x_{k+1})})
\end{aligned}
$$

Let $\alpha^2 L = \frac{1-\tau}{\tau}\alpha$

Both components telescope!

Linear Coupling

- Summing over $0, \ldots, T-1$, with $\bar{x} = \frac{1}{T}\sum_i x_i$

$$f(\bar{x}) - f(u) \leq \frac{\alpha L}{T}(f(y_0) - f(y_T)) + \frac{V_{z_0}(u)}{\alpha T}$$

Linear Coupling

- Summing over $0,\ldots,T-1$, with $\bar{x} = \frac{1}{T}\sum_i x_i$

$$f(\bar{x}) - f(u) \leq \frac{\alpha L}{T}(f(y_0) - f(y_T)) + \frac{V_{z_0}(u)}{\alpha T}$$

- If $f(y_0) - \mathrm{OPT} \leq d$, diameter $V_{z_0}(u) \leq \Theta$,
  $\alpha = \sqrt{\frac{\Theta}{Ld}}, T = 4\sqrt{\frac{L\Theta}{d}}$

$$f(\bar{x}) - f(u) \leq \frac{\alpha L d + \Theta/\alpha}{T} \leq \frac{d}{2}$$

Linear Coupling

- Summing over $0, \ldots, T-1$, with $\bar{x} = \frac{1}{T}\sum_i x_i$

$$f(\bar{x}) - f(u) \leq \frac{\alpha L}{T}(f(y_0) - f(y_T)) + \frac{V_{z_0}(u)}{\alpha T}$$

- If $f(y_0) - \mathrm{OPT} \leq d$, diameter $V_{z_0}(u) \leq \Theta$,
$\alpha = \sqrt{\frac{\Theta}{Ld}}, T = 4\sqrt{\frac{L\Theta}{d}}$

$$f(\bar{x}) - f(u) \leq \frac{\alpha L d + \Theta/\alpha}{T} \leq \frac{d}{2}$$

- In $T = 4\sqrt{\frac{L\Theta}{d}}$ iterations,

$$f(x_0) - \mathrm{OPT} \leq d \quad \rightarrow \quad f(\bar{x}) - \mathrm{OPT} \leq \frac{d}{2}$$

To get ε-approximation:

$$T = O(\sqrt{\frac{L\Theta}{\varepsilon}} + \sqrt{\frac{L\Theta}{2\varepsilon}} + \ldots) = O(\sqrt{\frac{L\Theta}{\varepsilon}})$$

Linear Coupling

- With $\alpha_k = \frac{k+1}{2L}$, can remove phases, and have $f(y_T) - f(u) \leq \varepsilon$ after $T = O(\sqrt{\frac{L\Theta}{\varepsilon}})$ iterations.
  Almost the same as Nesterov's.

- GD: $O(\frac{LR^2}{2})$ v.s. MD: $O(\frac{\rho^2\Theta}{\varepsilon^2})$ v.s. AGD: $O(\sqrt{\frac{L\Theta}{\varepsilon}})$