Convex Optimization

Prof. Nati Srebro

Lecture 17:
Proximal Methods
Mirror Descent

Bubeck Chapters 4 and 6; Ben Tal and Nemirovski Chapter 5

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ s.t. \quad & \|x\|_2 \leq 1 \end{aligned}$$

Assumption: $\|\nabla f(x)\|_2 \leq L$

We want to show:

For any method A that uses a 1st order oracle, there exists a function $f(\cdot)$ that satisfies the assumption (convex and L-Lipschitz) s.t. A requires $\geq \frac{L^2}{\epsilon^2}$ oracle accesses in order to find an ϵ suboptimal point.

Recall: we consider the answer returned by the algorithm as a "query", and play a game to construct a function such that all queries are at not-ϵ-suboptimal points.

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) \\ s.t. \quad & \|x\|_2 \le 1 \end{aligned}$$

Assumption: $\|\nabla f(x)\|_2 \le L$

- Consider very high dim, $n > 2\frac{L^2}{\epsilon^2}$ (otherwise can use center-of-mass)
- Construct a "hard" function of the form $f(x) = L \cdot \max_i \langle v_i, x\rangle$

  ➔ $\nabla f(x) = Lv_j$ s.t. $j = \arg\max_i \langle v_i, x\rangle$
- Given algorithm A, choose v_i adversarily by simulating A:

  On query $x^{(k)}$, pick $v_k \perp x^{(1)}, \dots, x^{(k)}, v_1, \dots, v_{k-1}$, $\|v_k\| = 1$

  answer using $f^{(k)}(x) = L \cdot \max_{i=1..k} \langle v_i, x\rangle$

Claim: $f^{(k)}(x)$ are convex and L-Lipschitz

Claim: $f^{(k)}\left(x^{(k)}\right) \ge 0$

Claim: For all $i \le k$, answer on $x^{(i)}$ using $f^{(i)}$ also valid for $f^{(k)}$

Proof: For $i < j \le k$, $\left\langle v_j, x^{(i)}\right\rangle = 0 \le f^{(i)}\left(x^{(i)}\right)$ so can ignore in max and argmax

Consider $x^* = \frac{-1}{\sqrt{T}}\sum_{i=1}^{T} v_i$, then $\|x^*\| = 1$ and $f^{(T)}(x^*) = \frac{-L}{\sqrt{T}}$

➔ $\min_{i=1..T} f^{(T)}\left(x^{(i)}\right) \ge 0 \ge f^{(T)}(x^*) + \frac{L}{\sqrt{T}}$

Conclusion: A requires $\ge L^2/\epsilon^2$ queries to find ϵ-suboptimal of f

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ s.t. \quad & \|x\|_2 \leq 1 \end{aligned}$$

Assumption: $|f(x)| \leq B$

- Same construction shows need $\Omega(n)$ queries to find optimum
- Combine with $\log \frac{1}{\epsilon}$ construction in one dimension to establish $\Omega\left(n \log \frac{B}{\epsilon}\right)$ lower bound.

Convex Optimization

Prof. Nati Srebro

Lecture 17:
Proximal Methods
Mirror Descent

$$x^{(k+1)} \leftarrow \arg\min_{x \in \mathcal{X}} f^{(k)}(x) + \frac{\lambda}{2}\left\|x - x^{(k)}\right\|^2$$

- Gradient Descent:

$$f^{(k)}(x) = f\left(x^{(k)}\right) + \langle \nabla f\left(x^{(k)}\right), x - x^{(k)} \rangle$$
$$= \langle \nabla f\left(x^{(k)}\right), x \rangle + const$$

- Trust-Region Newton:

$$f^{(k)}(x) = f\left(x^{(k)}\right) + \langle \nabla f\left(x^{(k)}\right), x - x^{(k)} \rangle + \frac{1}{2}\left(x - x^{(k)}\right)^{\top} \nabla^2 f\left(x^{(k)}\right)\left(x - x^{(k)}\right)$$

- Stochastic Gradient Descent:

$$f^{(k)}(x) = \langle g^{(k)}, x \rangle + const$$
$$\mathbb{E}\left[g^{(k)}\right] = \nabla f^{(k)}\left(x^{(k)}\right)$$

- “Aggressive” Stochastic Descent:

$$\mathbb{E}\left[f^{(k)}\right] = f$$

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$$

e.g. $f_i(x) = loss(\langle a_i, x\rangle - b_i)$

- Gradient Descent: $f^{(k)}(x) = \langle \frac{1}{m}\sum_{i=1}^{m} \nabla f_i\left(x^{(k)}\right), x\rangle$
  $= \langle \frac{1}{m}\sum_{i=1}^{m} loss'\left(\langle a_i, x^{(k)}\rangle - b\right)a_i, x\rangle$
  - Cost per iteration: m ∇f_i evals ➔ $O(mn)$
    + $\arg\min_x \langle v, x\rangle + \left\|x - x^{(k)}\right\|^2$ ➔ $O(n)$

- SGD: $f^{(k)}(x) = \langle \nabla f_{i^{(k)}}\left(x^{(k)}\right), x\rangle$ for random $i^{(k)}$
  - Cost per iteration: single ∇f_i eval ➔ $O(n)$
    + $\arg\min_x \langle v, x\rangle + \left\|x - x^{(k)}\right\|^2$ ➔ $O(n)$

- Aggressive Stochastic: $f^{(k)}(x) = f_{i^{(k)}}(x)$ for random $i^{(k)}$
  - Each iteration: $\arg\min_x f_i(x) + \left\|x - x^{(k)}\right\|^2$
  - If $loss()$ piecewise linear or quadratic ➔ $O(n)$

Partial Linearization

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x) + h(x)$$

e.g. $h(x) = \|x\|_1$

- Partially linearized stochastic:

$$f^{(k)}(x) = \langle \nabla f_{i(k)}\left(x^{(k)}\right), x\rangle + h(x)$$

Partial Linearization

$$f(x) = g(x) + h(x)$$

- Use $f^{(k)}(x) = \langle \nabla g(x^{(k)}), x \rangle + h(x)$, i.e.

$$x^{(k+1)} \leftarrow \arg\min_x \langle \nabla g(x^{(k)}), x \rangle + h(x) + \frac{\lambda}{2} \|x - x^{(k)}\|^2$$

- Required oracles:
  - 1st order oracle for g: $x \rightarrow g(x), \nabla g(x)$
  - Prox oracle for h: $y, \lambda \rightarrow \arg\min_x h(x) + \frac{\lambda}{2} \|x - y\|^2$
- If $g(x)$ is M-smooth and $h(x)$ non-smooth but has prox oracle:
  - Only $O\left(M\|x^*\|^2/\epsilon\right)$ iterations
  - Can use Nesterov acceleration to reduce to $O\left(\sqrt{M\|x^*\|^2/\epsilon}\right)$
- Non-linearized $h(x)$ need not be smooth, Lipschitz, or even bounded
- Generalizes projected gradient descent
  - $h_{\mathcal{X}}(x) = 0 \; if \; x \in \mathcal{X}, \infty \; otherwise$
  - $x^{(k+1)} \leftarrow \arg\min_{x \in \mathcal{X}} \langle \nabla g(x^{(k)}), x \rangle + \frac{\lambda}{2} \|x - x^{(k)}\|^2 = \Pi_{\mathcal{X}}\left(x^{(k)} - \frac{1}{\lambda} \nabla g(x^{(k)})\right)$

Full vs Partial Linearization: Example

$$f(x) = g(x) + \|x\|_1$$

smooth

- $\|x\|_1$ penalty often added to encourage sparsity in the solution
- Since $[-1,1] \in \partial|x|$, we will not move away from $x[i] = 0$ if $\left|\frac{\partial g(x)}{\partial x[i]}\right| \leq 1$
- Using (fully linearized) gradient descent:
  - Non-smooth objective, slow $1/\epsilon^2$ convergence
  - Iterates never sparse (nothing to encourage sparseness of iterates)
- Option 1: Cast as constrained optimization

$$\min_{x \in \mathbb{R}^n, t \in \mathbb{R}^n} g(x) + \sum_i t_i \quad s.t. \quad -t_i \leq x_i \leq t$$

  - $O(n^{3.5} \log {}^1/_\epsilon)$ runtime—bad dependence on n
  - If using Interior-Point methods, iterates never sparse
- Option 2: Partial linearization
  - Faster convergence, can use acceleration: $O\left({}^{n\|x^*\|}/_{\sqrt{\epsilon}}\right)$ runtime
  - Encourage sparsity of iterates

Bundle Methods

$$\arg\min_x f(x)$$

- Use:

$$f^{(k)} = \max_{i=0..k} f\left(x^{(i)}\right) + \langle \nabla f\left(x^{(i)}\right), x - x^{(i)} \rangle$$

- Recall that for convex $f(x)$, each (sub)gradient provides a lower bound, so their max is also a valid lower bound

Other Geometries

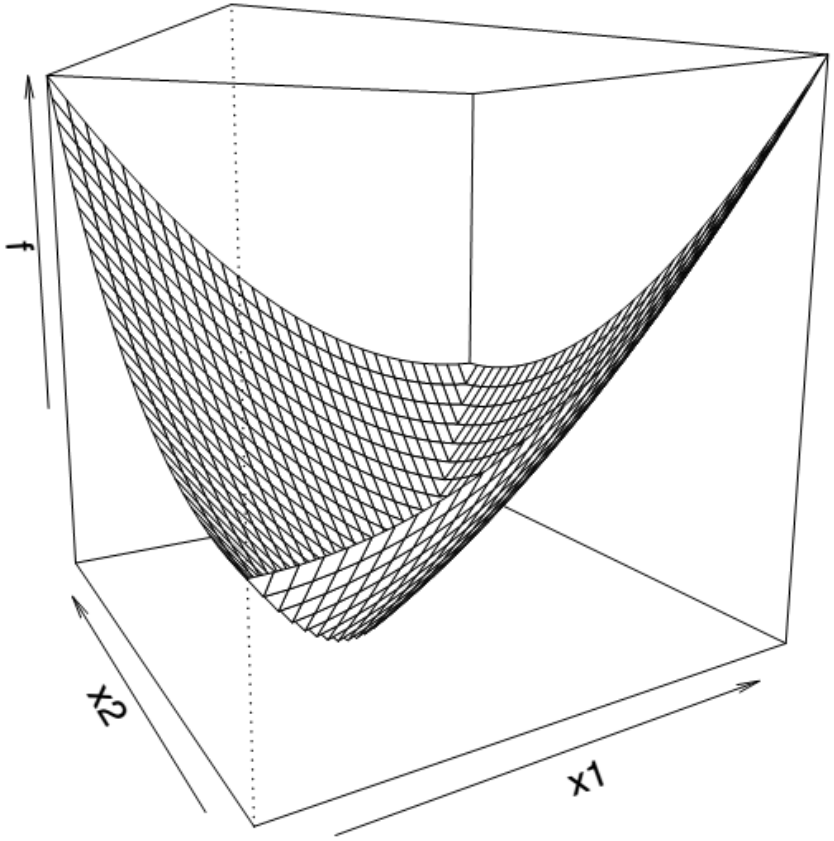
$$x^{(k+1)} \leftarrow \arg\min_{x \in \mathcal{X}} f^{(k)}(x) + \lambda D(x; x^{(k)})$$

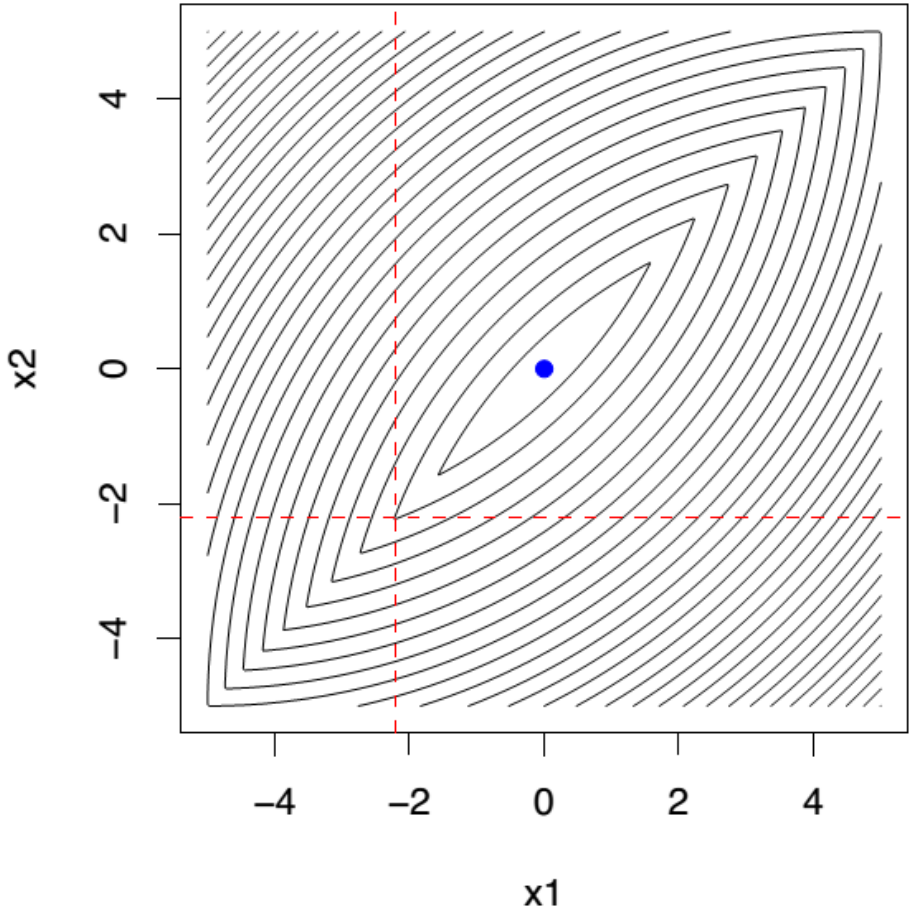
- So far: $D(x; y) = \frac{1}{2}\|x - y\|_2^2$
- Also other quadratic norms: $D(x; y) = \frac{1}{2}\|x - y\|_Q^2 = \frac{1}{2}(x - y)^\top Q(x - y)$
  - Equivalent to change of variables to $\tilde{x} = Q^{1/2}x$
- Steepest Descent: $D(x; y) = \|x - y\|^2$
  - With $f^{(k)}(x) = \langle \nabla f(x^{(k)}), x \rangle$ and $\mathcal{X} = \mathbb{R}^n$
  - Alternate view: infinitesimally as $\lambda \to \infty$

$$\Delta x^{(k)} = \lambda \cdot \left(\lim_{\lambda \to \infty} \arg\min_{\Delta x} f(x^{(k)} + \Delta x) + \lambda \|\Delta x\|^2 \right)$$

  - E.g. $\|\cdot\|_1$ ➔ coordinate descent; $\|\cdot\|_\infty$ ➔ $\Delta x = sign(\nabla f)$
  - Convergence properties could be bad

Coordinate descent can get stuck with non-smooth objectives:





Goal: From Potential Function to Divergence

- Recall Gradient Descent Guarantee:

$|f(x) - f(y)| \leq L\|x - y\|_2$

$$k = O\left(\frac{L^2 \|x^*\|_2^2}{\epsilon^2}\right)$$

$f(x + \Delta x) \leq f(x) + \langle \nabla f(x), \Delta x \rangle + \frac{M}{2}\|\Delta x\|_2^2$

$$k = O\left(\frac{M \|x^*\|_2^2}{\epsilon}\right)$$

$$k = O\left(\frac{M}{\mu} \log \frac{\epsilon_0}{\epsilon}\right)$$

$f(x) + \langle \nabla f(x), \Delta x \rangle + \frac{\mu}{2}\|\Delta x\|_2^2 \leq f(x + \Delta x)$

- Dependence on $\Psi(x) = \frac{1}{2}\|x\|^2$
- **Can we get dependence on different $\Psi(x)$?**

Intuition for Constructing Divergence: Revisiting Gradient Descent

- Consider:

$$x^{(k+1)} \overset{\text{def}}{=} \arg\min_{x} \sum_{i=1}^{k} \langle \nabla f(x^{(k)}), x \rangle + \lambda \Psi(x)$$

- We have that:

$$0 = \sum_{i=1}^{k} \nabla f(x^{(k)}) + \lambda \nabla \Psi(x^{(k+1)})$$

$$\Rightarrow x^{(k+1)} = \nabla \Psi^{-1}\left(-\frac{1}{\lambda} \sum_{i=1}^{k} \nabla f(x^{(k)})\right)$$

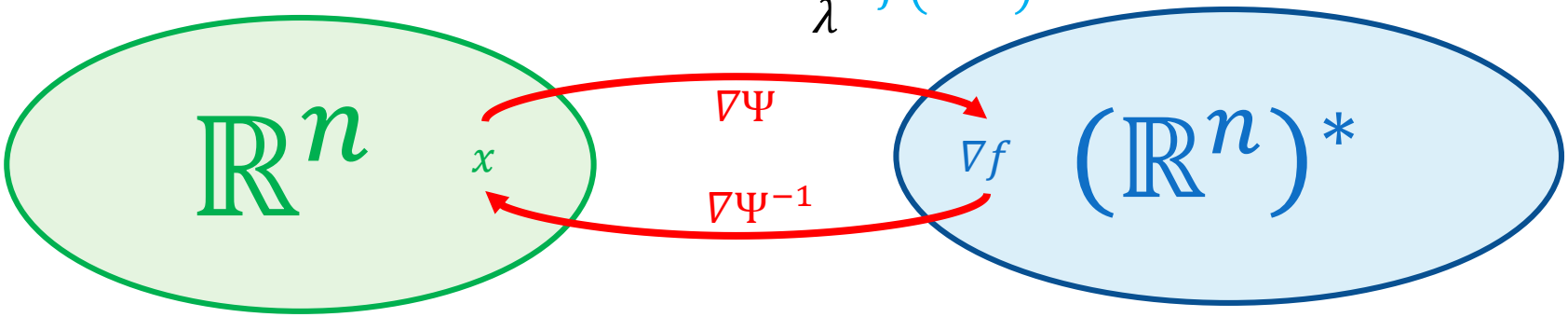
$$= \nabla \Psi^{-1}\left(\nabla \Psi(x^{(k)}) - \frac{1}{\lambda} \nabla f(x^{(k)})\right)$$

Dual Averaging

Mirror Descent

- For $\Psi(x) = \frac{1}{2}\|x\|^2$ we have $\nabla \Psi(x) = x$ and so:

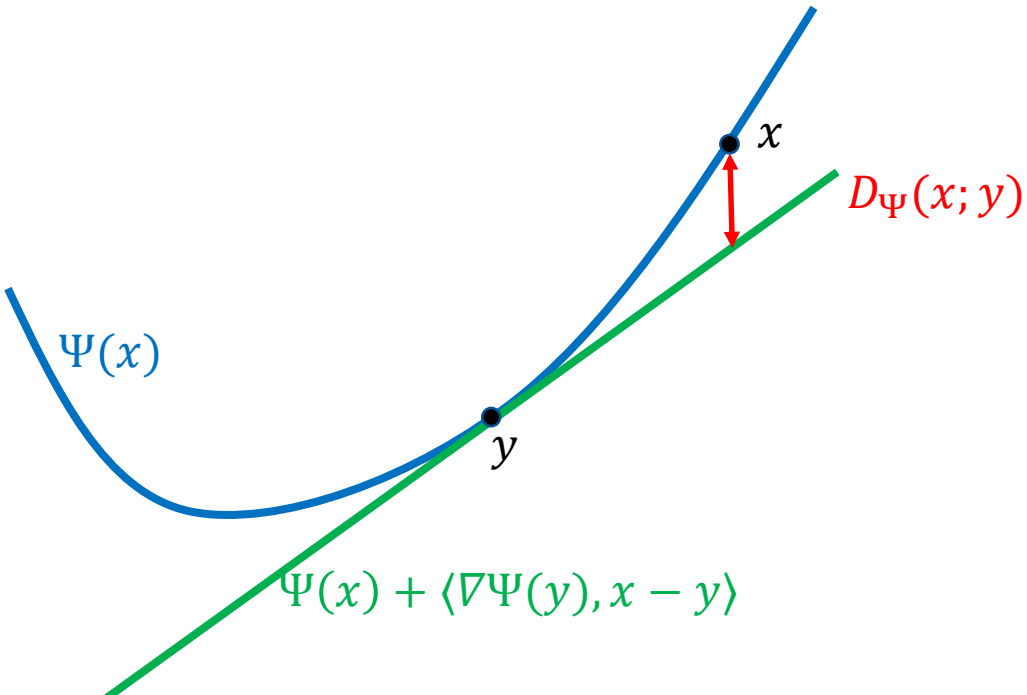
$$x^{(k+1)} = x^{(k)} - \frac{1}{\lambda} \nabla f(x^{(k)})$$



Bergman Divergence

$$D_\Psi(x;y) = \Psi(x) - (\Psi(y) + \langle \nabla\Psi(y), x - y\rangle)$$

- Ψ convex $\Leftrightarrow D_\Psi(x;y) \geq 0$
- Ψ strictly convex ➔ $D_\Psi(x;y) = 0$ only for $x = y$
- Ψ α-strongly convex w.r.t. $\|x\|$ ➔ $D_\Psi(x;y) \geq \frac{\alpha}{2}\|x - y\|^2$



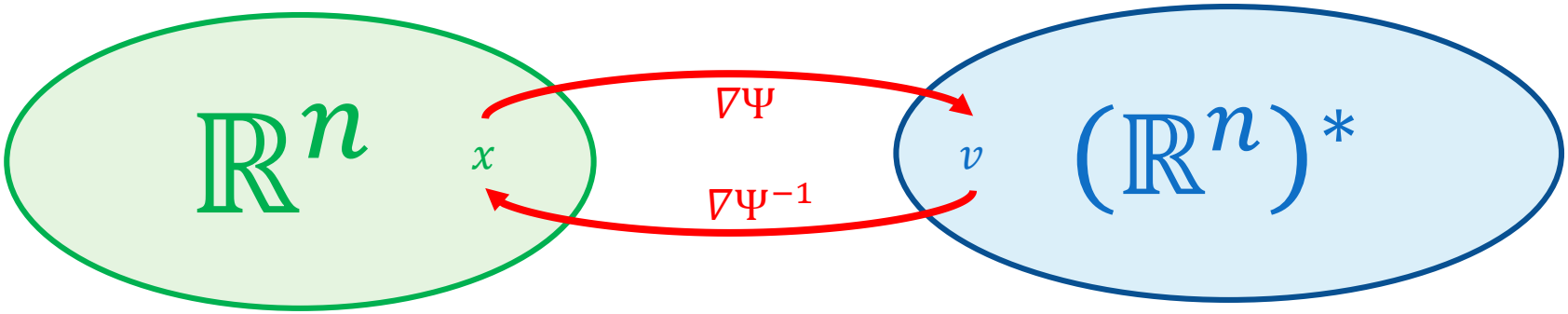
Bergman Divergence

$$D_\Psi(x;y) = \Psi(x) - (\Psi(y) + \langle \nabla\Psi(y), x - y\rangle)$$

- Ψ convex $\Leftrightarrow D_\Psi(x;y) \geq 0$
- Ψ strictly convex ➔ $D_\Psi(x;y) = 0$ only for $x = y$
- Ψ α-strongly convex w.r.t. $\|x\|$ ➔ $D_\Psi(x;y) \geq \frac{\alpha}{2}\|x - y\|^2$
- Claim:

$$\arg\min_x \langle v, x\rangle + D_\Psi(x;y) = \nabla\Psi^{-1}(\nabla\Psi(y) - v)$$

Proof: $0 = v + \nabla\Psi(x) - \nabla\Psi(y)$



Mirror Descent (Generic)

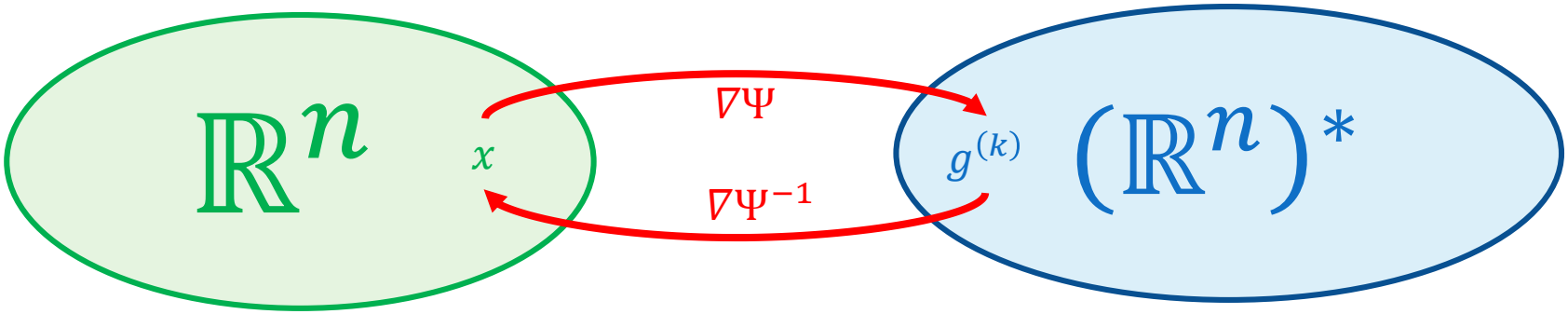
$$x^{(k+1)} \leftarrow \arg\min_{x \in \mathcal{X}} f^{(k)}(x) + \lambda^{(k)} D_\Psi(x; x^{(k)})$$

$$D_\Psi(x; y) = \Psi(x) - (\Psi(y) + \langle \nabla\Psi(y), x - y \rangle)$$

- For $f^{(k)}(x) = \langle g^{(k)}, x \rangle$:

$$x^{(k+1)} = \Pi_{\mathcal{X}}^{\Psi}\left(\nabla\Psi^{-1}\left(\nabla\Psi\left(x^{(k)}\right) - \frac{1}{\lambda^{(k)}} g^{(k)}\right)\right)$$

where $\Pi_{\mathcal{X}}^{\Psi}(y) = \arg\min_{x \in \mathcal{X}} D_\Psi(x; y)$



Mirror Descent—Analysis

$$x^{(k+1)} \leftarrow \arg\min_{x \in \mathcal{X}} \langle \nabla f\left(x^{(k)}\right), x \rangle + \lambda D_\Psi(x; x^{(k)})$$

$$D_\Psi(x; y) = \Psi(x) - (\Psi(y) + \langle \nabla \Psi(y), x - y \rangle)$$

- Method defined in terms of Ψ
- Requires:
  - 1st order oracle $x \to f(x), \nabla f(x)$
  - Prox operator for Ψ: $v \to \arg\min_{x \in \mathcal{X}} \langle v, x \rangle + \Psi(x)$

    or

  - $x \to \nabla \Psi^{-1}(x)$ **and** $v \to \nabla \Psi(v)$ **and** $x \to \Pi_{\mathcal{X}}^{\Psi}(x)$
- Analysis relies also on some norm $\|x\|$
  - Ψ α-strongly convex w.r.t. $\|x\|$, i.e. $D_\Psi(x + \Delta x; x) \geq \frac{\alpha}{2} \|\Delta x\|^2$
  - Objective f is L-Lipschitz w.r.t $\|x\|$, i.e. $|f(x) - f(y)| \leq L\|x - y\|$
    equivalently: $\|\nabla f(x)\|_* \leq L$

Mirror Descent—Analysis

$$x^{(k+1)} \leftarrow \arg\min_{x \in \mathcal{X}} \langle \nabla f\left(x^{(k)}\right), x \rangle + \lambda D_{\Psi}(x; x^{(k)})$$

If:

- Ψ α-strongly convex w.r.t. $\|x\|$, i.e. $D_{\Psi}(x + \Delta x; x) \geq \frac{\alpha}{2} \|\Delta x\|^2$
- Objective f is L-Lipschitz w.r.t $\|x\|$, i.e. $|f(x) - f(y)| \leq L\|x - y\|$
  equivalently: $\|\nabla f(x)\|_* \leq L$

Number of iterations to ensure ϵ suboptimality:

$$k = O\left(\frac{L^2 \Psi(x^*)}{\alpha\, \epsilon^2}\right)$$

- Also with stochastic gradients
- Similarly generalizes analysis for smooth and strongly convex, including acceleration
- Smoothness and strong convexity also w.r.t. $\|x\|$

Example: $\|x\|_2$

- $\Psi(x) = \frac{1}{2}\|x\|_2^2$ is 1-strongly convex
- $\nabla\Psi(x) = x, \nabla\Psi^{-1}(\nu) = \nu$
- $D_\Psi(x'; x) = \Psi(x') - (\Psi(x) + \langle\nabla\Psi(x), x' - x\rangle)$

$$= \frac{1}{2}\|x'\|_2^2 - \left(\frac{1}{2}\|x\|_2^2 + \langle x, x' - x\rangle\right) + \|x\|_2^2 - \langle x, x\rangle$$

$$= \frac{1}{2}\|x - x'\|_2^2$$

Example: $\|x\|_Q = \sqrt{x^\top Q x}$

- $\Psi(x) = \frac{1}{2} x^\top Q x$ is 1-strongly convex w.r.t $\|x\|_Q$
- $\nabla\Psi(x) = Qx, \nabla\Psi^{-1}(\nu) = Q^{-1}\nu$
- $D_\Psi(x'; x) = \frac{1}{2}\|x - x'\|_Q^2$

- Mirror Descent step: $x^{(k+1)} = x^{(k)} - \frac{1}{\lambda} Q^{-1} \nabla f\left(x^{(k)}\right)$

- $\|\nu\|_* = \nu^\top Q^{-1} \nu$
- Iterations:

$$O\left(\frac{\left(x^{*\top} Q x^*\right)\left(\nabla f^\top Q^{-1} \nabla f\right)}{\epsilon^2}\right)$$

Example: $\|x\|_p$

- $\Psi(x) = \frac{1}{2}\|x\|_p^2$ is $(p-1)$-strongly convex w.r.t. $\|x\|_p$
- $\nabla\Psi(x) = \|x\|_p^{2-p}|x[i]|^{p-1}sign(x[i])$
- $\nabla\Psi^{-1}(v) = \frac{|v[i]|^{q-1}sign(v[i])}{\|v\|_q^{q-2}}$, where $\frac{1}{p} + \frac{1}{q} = 1$
- Iterations: $O\left(\frac{\|x^*\|_p^2\|\nabla f\|_q^2}{(p-1)\epsilon^2}\right)$
- Explodes as $p \to 1$

- What about $\|x\|_1$?
  - Option 1: use $q = \log n$
  - Option 2: entropy potential

$\mathcal{X} = \{x \mid x \geq 0, \|x\|_1 = 1\}$

- $\Psi(x) = \sum_i x[i] \log \frac{x[i]}{1/n} = \log n + \sum_i x[i] \log x[i]$
- Claim: $\Psi(x)$ is 1-strongly convex w.r.t. $\|x\|_1$ on $\mathcal{X}$
- For $\|x\|_1 = 1$: $0 \leq \Psi(x) \leq \log n$
- $\nabla\Psi(x)[i] = \log x[i] + 1$
- $\nabla\Psi^{-1}(v)[i] = e^{v[i]-1}$
- Exponentiated Gradient (Multiplicative Updates):

$$x^{(k+1)} = \arg\min_{x \in \mathcal{X}} \langle g^{(k)}, x \rangle + \lambda\Psi(x)$$

$$\rightarrow x^{(k+1)}[i] \propto e^{-\frac{1}{\lambda} g^{(k)}[i]} x^{(k)}$$

- Iterations: $O\left(\frac{\|\nabla f(x^{(k)})\|_\infty^2 \log n}{\epsilon^2}\right)$

Mirror Descent (Generic)

$$x^{(k+1)} \leftarrow \arg\min_{x \in \mathcal{X}} f^{(k)}(x) + \lambda^{(k)} D_\Psi(x; x^{(k)})$$

$$D_\Psi(x; y) = \Psi(x) - (\Psi(y) + \langle \nabla\Psi(y), x - y \rangle)$$

- For $f^{(k)}(x) = \langle g^{(k)}, x \rangle$:

$$x^{(k+1)} = \Pi_{\mathcal{X}}^{\Psi}\left(\nabla\Psi^{-1}\left(\nabla\Psi\left(x^{(k)}\right) - \frac{1}{\lambda^{(k)}} g^{(k)}\right)\right)$$

where $\Pi_{\mathcal{X}}^{\Psi}(y) = \arg\min_{x \in \mathcal{X}} D_\Psi(x; y)$

- Can use all other $f^{(k)}$ discussed (stochastic, aggressive, partial linearization, 2nd order, bundle, …) and also acceleration
- Gradient-Descent-like ($\Psi(x) = \frac{1}{2}\|x\|_2^2$) guarantees generally carry over, with dependence on strong convexity of $\Psi(x)$ w.r.t relevant norm $\|x\|$, and on properties of objective w.r.t. same $\|x\|$