Mirror Descent
based on lecture notes by Yuxin Chen (Princeton)

Themis       Kurt
Gradient Descent for Function Minimization

\[ x^{t+1} = x^t - \eta_t \nabla f(x^t) \quad \text{small step in direction of the negative gradient} \]

\[ = \arg \min_{x} \left\{ f(x^t) + \langle \nabla f(x^t), x - x^t \rangle + \frac{1}{2\eta_t} \|x - x^t\|^2 \right\}. \]

- We approximate \( f \) by a quadratic function that passes through \((x^t, f(x^t))\) and has the same gradient as \( f \) at \( x^t \).
- We move to the minimizer of the quadratic function; \( x^{t+1} \) is the solution of \( \nabla f(x^t) + \frac{1}{\eta_t} (x - x^t) = 0 \).
- At \( x^{t+1} \), the gradient of the quadratic term is \( -\nabla f(x^t) \).
Gradient Descent

We are also interested in constrained optimization: \( C \) is a convex subset of \( \mathbb{R}^n \).

\[
x^{t+1} = \arg \min_{x \in C} \left\{ f(x^t) + \langle \nabla f(x^t), x - x^t \rangle + \frac{1}{2\eta_t} \| x - x^t \|_2^2 \right\}.
\]

Why are we approximating by a homogeneous quadratic function?

Aren’t there other (better?) choices?
Clearly, there are better Choices sometimes

Assume \( f \) is a quadratic function, i.e.,
\[
f(x) = \frac{1}{2} (x - x^t)^T Q (x - x^t)
\] with \( Q \) positive semidefinite.

Then we should clearly approximate with the function itself. Iteration becomes

\[
x^{t+1} = \arg \min_x \left \{ f(x^t) + \langle \nabla f(x^t), x - x^t \rangle + \frac{1}{2 \eta_t} (x - x^t)^T Q (x - x^t) \right \}
\]
\[
= x^t - \eta_t Q^{-1} \nabla f(x^t)
\]

Note that at \( x^{t+1} \):
\[
- \nabla f(x^t) = \frac{1}{\eta_t} Q (x^{t+1} - x^t).
\]

With \( \eta_t = 1 \), we would reach the minimum in one step.

If \( Q \) is a diagonal matrix with \( \kappa = \frac{\max_i Q_{ii}}{\min_i Q_{ii}} \gg 1 \), GD is slow: \( \kappa \log(1/\varepsilon) \) iterations.

Alejandro’s talk: Newton iteration, \( \alpha H \prec A \prec \beta H \).
Mirror descent: choose proximity term to fit problem geometry

Nemirowski & Yudin, 1983

- local curvature of $f$
- geometry of the constraint set $C$
- computation of $x^{t+1}$ is efficient.
Mirror Descent

Replace the quadratic term by a “distance function” \( D_\varphi \).

\[
x^{t+1} = \arg \min_{x \in C} \left\{ f(x^t) + \langle \nabla f(x^t), x - x^t \rangle + \frac{1}{\eta_t} D_\varphi(x, x^t) \right\}
\]

\[
D_\varphi(x, z) = \varphi(x) - (\varphi(z) + \langle \nabla \varphi(z), (x - z) \rangle).
\]

- \( D_\varphi(x, z) \) is distance from \( z \) to \( x \) with respect to \( \varphi \); \( \varphi \) is strongly convex and differentiable.
- Bregman divergence; Lev Bregman, 1967.
- at \( x^{t+1} \) gradient of \( \frac{1}{\eta_t} D_\varphi(x, x^t) \) is equal to \(-\nabla f(x^t)\).
- more generally,

\[
x^{t+1} = \arg \min_{x \in C} \left\{ f(x^t) + \langle g^t, x - x^t \rangle + \frac{1}{\eta_t} D_\varphi(x, x^t) \right\}
\]

with \( g^t \) a subgradient of \( f \) at \( x^t \); \( g^t \in \partial f(x^t) \).
Properties of Bregman Divergence

\[ D_\varphi(x, z) = \varphi(x) - (\varphi(z) + \langle \nabla \varphi(z), (x - z) \rangle) \].

- distance from \( z \) to \( x \) with respect to \( \varphi \); \( \varphi \) is strongly convex and differentiable.
- \( D_\varphi(x, z) \geq 0 \) and equal to 0 only if \( x = z \).
- \( \nabla_x D_\varphi(x, z) = \nabla \varphi(x) - \nabla \varphi(z) \).
- in general \( D_\varphi(x, z) \neq D_\varphi(z, x) \).
- convex in \( x \), in general not convex in \( z \).
- if \( Q \succ 0 \) and \( \varphi(x) = x^T Q x \), then \( D_\varphi(x, z) = \frac{1}{2} (x - z)^T Q (x - z) \).
  So gradient descent is a special case (even with non-homogeneous quadratic function).
Kullback-Leibler Divergence

- directed distance between two probability distributions; introduced in 1951.

\[ \varphi(x) = \sum_i x_i \ln x_i \] negative entropy

- for \( x, z \in \Delta = \left\{ x \in \mathbb{R}^n_{\geq 0}; \sum_i x_i = 1 \right\} \) (probability simplex)

\[ \text{KL}(x \parallel z) = D_\varphi(x, z) = \sum_i x_i \ln(x_i/z_i). \]

- Proof: Since \( (\nabla \varphi(x))_i = \ln x_i + 1 \)

\[ D_\varphi(x, z) = \varphi(x) - (\varphi(z) + \nabla \varphi(z)(x - z)) \]

\[ = \sum_i x_i \ln x_i - \sum_i z_i \ln z_i - \sum_i (\ln z_i + 1)(x_i - z_i) \]

\[ = \sum_i x_i \ln(x_i/z_i) - \sum_i x_i + \sum_i z_i \]

\[ = \sum_i x_i \ln(x_i/z_i). \]
The Update Rule for Mirror Descent with KL Divergence in Probability Simplex

\[ x^{t+1} = \arg \min_{x \in \Delta} \left\{ f(x^t) + \langle \nabla f(x^t), x - x^t \rangle + \frac{1}{\eta_t} \text{KL}(x \| x^t) \right\} \]

\[ \text{KL}(x \| x^t) = \sum_i x_i \ln(x_i / x_i^t) \]

At \( x^{t+1} \), gradient of objective must be parallel to normal of \( \Delta \) (the all-ones vector), i.e., there must be an \( \alpha \) such that for all \( i \) with \( x_i^{t+1} \notin \{ 0, 1 \} \)

\[ (\nabla f(x^t))_i + \frac{1}{\eta_t} \left[ \ln(x_i^{t+1} / x_i^t) + x_i^{t+1} \cdot x_i^t / x_i^{t+1} \cdot 1 / x_i^t \right] = \alpha \cdot 1 \]

and hence \( x_i^{t+1} / x_i^t = \exp(-\eta_t(\nabla f(x^t))_i + \eta_t \alpha - 1) \) or

\[ x_i^{t+1} = x_i^t \exp(-\eta_t(\nabla f(x^t))_i) / C \quad \text{for some constant } C. \]

Since \( x^{t+1} \in \Delta, C = \sum_i x_i^t \exp(-\eta_t(\nabla f(x^t))_i). \)
Alternative View of Mirror Descent.

- **Bregman projection of** $x$ **onto** $C$

  $$\mathcal{P}_{C,\varphi}(x) = \arg\min_{z \in C} D_\varphi(z, x)$$

  the point $z \in C$ closest to $x$ with respect to $D_\varphi$.

- **Unconstrained mirror descent**

  $$x^{t+1} = \arg\min_x \left\{ f(x^t) + \langle \nabla f(x^t), x - x^t \rangle + \frac{1}{\eta_t} D_\varphi(x, x^t) \right\}$$

  $$\nabla \varphi(x^{t+1}) = \nabla \varphi(x^t) - \eta_t \nabla f(x^t)$$

- **Alternative view of constrained mirror descent**

  $$\nabla \varphi(y^{t+1}) = \nabla \varphi(x^t) - \eta_t \nabla f(x^t)$$

  $$x^{t+1} = \mathcal{P}_{C,\varphi}(y^{t+1}) = \arg\min_x D_\varphi(x, y^{t+1})$$

  Unconstrained step followed by Bregman projection onto $C$. 
Proof of Equivalence

\[ x^{t+1} = \arg \min_{x \in C} \left\{ f(x^t) + \langle \nabla f(x^t), x - x^t \rangle + \frac{1}{\eta_t} D_\varphi(x, x^t) \right\} \]

Optimality condition: Negative gradient of \{ \ldots \} in normal cone of \mathcal{C} at \(x^{t+1}\).

\[ - \left( \nabla f(x^t) + \frac{1}{\eta_t} (\nabla \varphi(x^{t+1}) - \nabla \varphi(x^t)) \right) \in \mathcal{N}_\mathcal{C}(x^{t+1}). \]

\[ \nabla \varphi(y^{t+1}) = \nabla \varphi(x^t) - \eta_t \nabla f(x^t) \]

\[ x^{t+1} = \mathcal{P}_{\mathcal{C}, \varphi}(y^{t+1}) = \arg \min_{x \in \mathcal{C}} D_\varphi(x, y^{t+1}) \]

Optimality condition: negative gradient of \( D_\varphi(x, y^{t+1}) \) in normal cone at \(x^{t+1}\).

\[ - \left( \nabla \varphi(x^{t+1}) - \nabla \varphi(y^{t+1}) \right) \in \mathcal{N}_\mathcal{C}(x^{t+1}). \]

Optimality conditions are identical.
Assume $C = \mathbb{R}^n$ for simplicity. Then

$$x^{t+1} = \nabla \varphi^*\left((\nabla \varphi(x^t) - \eta_t \nabla f(x^t))\right),$$

where $\varphi^*$ is the Fenchel-conjugate of $\varphi$.

$$\varphi^*(y) = \sup_z [\langle z, x \rangle - \varphi(z)]$$
Convergence of Mirror Descent to $\min_{x \in C} f(x)$

$\|\|$ is a norm

Assume $f$ is convex and $L$-Lipschitz.

Assume $\varphi$ is $\rho$-strongly convex wrt. $\|\|$.

Run mirror descent for $t$ steps starting at $x^0$: $x^0, x^1, \ldots, x^t$.

Let $f^{\text{best},t} = \min_{0 \leq i \leq t} f(x^i)$ and $R = \sup_{x \in C} D_\varphi(x, x^0)$.

Then

$$f^{\text{best},t} - f^{\text{opt}} \leq \frac{R + \frac{L}{2\rho} \sum_{0 \leq k < t} \eta_k^2}{\sum_{0 \leq k < t} \eta_k}$$

$$= L \cdot \sqrt{\frac{2R}{\rho t}} \quad \text{with} \quad \eta_k = \frac{\sqrt{2\rho R}}{L \sqrt{t}}$$
- \( f \) is convex:

\[
f(y) \geq f(x) + \langle \nabla f(x)^T, y - x \rangle.
\]

- \( \varphi \) is \( \rho \)-strongly convex wrt. \( \| \| \), i.e.,

\[
\varphi(x) \geq \varphi(y) + \langle \nabla \varphi(y), x - y \rangle + \frac{\rho}{2} \| x - y \|^2.
\]

- \( f \) is \( L \)-Lipschitz:

\[
|f(x) - f(y)| \leq L \cdot \| x - y \|.
\]
Convergence of Mirror Descent to $\min_{x \in \mathcal{C}} f(x)$

$\|\|$ is a norm

Assume $f$ is convex and $L$-Lipschitz.

Assume $\varphi$ is $\rho$-strongly convex wrt. a norm $\|\|$.

Run mirror descent for $t$ steps starting at $x^0$: $x^0, x^1, \ldots, x^t$.

Let $f_{\text{best},t} = \min_{0 \leq i \leq t} f(x^i)$ and $R = \sup_{x \in \mathcal{C}} D_{\varphi}(x, x^0)$.

Then

$$f_{\text{best},t} - f_{\text{opt}} \leq \frac{R + \frac{L}{2\rho} \sum_{0 \leq k < t} \eta_k^2}{\sum_{0 \leq k < t} \eta_k}$$

$$= L \cdot \sqrt{\frac{2R}{\rho t}} \quad \text{with} \quad \eta_k = \frac{\sqrt{2\rho R}}{L \sqrt{t}}$$

max planck institut informatik 15
Gradient vs Mirror over the Probability Simplex

- $C = \Delta$ (probability simplex) and $x^0 = n^{-1}1$.
- $\varphi(x) = \frac{1}{2} \| x \|_2^2$ is 1-strongly convex w.r.t. $\| \|_2$.
- $R = \sup_{x \in \Delta} D_{\varphi}(x, x^0) \leq 1/2$ and $L_{f,2} = \sup_{x \in \Delta} \| \nabla f(x) \|_2$.
- Then
  $$f^{\text{best},t} - f^{\text{opt}} \leq L_{f,2} \cdot \frac{1}{\sqrt{t}}$$

- $\varphi(x) = \sum_i x_i \ln x_i$ is 1-strongly convex w.r.t. $\| \|_1$.
- $R = \sup_{x \in \Delta} KL(x \| x^0) = \sup_{x \in \Delta} \sum_i x_i \ln x_i - \sum_i x_i \ln \frac{1}{n} \leq 0 + \ln n$.
- $L_{f,\infty} = \sup_{x \in \Delta} \| \nabla f \|_\infty$.
- Then
  $$f^{\text{best},t} - f^{\text{opt}} \leq L_{f,\infty} \cdot \frac{1}{\sqrt{t}}$$

- Since $\| \|_\infty \leq \| \|_2 \leq \sqrt{n} \| \|_\infty$, MD is often much better.
Robust Regression (taken from Stanford EE364B)

- minimize $\|Ax - b\|_1 = \sum_{1 \leq i \leq m} |a_i^T x - b_i|$ subject to $x \in \Delta$.

- Subgradient of objective is $g = \sum_{1 \leq i \leq m} \text{sign}(a_i^T x - b_i) a_i$.

- Projected subgradient update ($\varphi(x) = \|x\|_2^2$) is:
  Let $y^{t+1} = x^t + \eta_t g^t$. Then $x^{t+1} = \arg\min_{x \in \Delta} \|x - y^{t+1}\|_2$.
  Let $z \in \mathbb{R}^n$ be the orthogonal projection of $y^{t+1}$ onto hyperplane $1^T z = 1$.
  Then $x_i^{t+1}$ = see drawing

- Mirror descent update ($\varphi(x) = \sum_i x_i \ln x_i$) is (see slide 9):
  \[
  x_i^{t+1} = \frac{x_i^t \exp(-\eta_t g_i^t)}{\sum_j x_j^t \exp(-\eta_t g_j^t)}.
  \]
Robust regression problem with \( a_i \sim N(0, I_{n \times n}) \) and 
\[ b_i = (a_{i,1} + a_{i,2})/2 + \varepsilon_i \]
where \( \varepsilon_i \sim N(0, 10^{-2}) \), \( m = 20 \), \( n = 3000 \)

solution is close to \( x_1 \approx 1/2, \ x_2 \approx 1/2 \).

What they call \( k \), we call \( t \).