# 13. Generalized distances and mirror descent

- Bregman distance
- properties
- Bregman proximal mapping
- mirror descent

# Motivation: proximal gradient method

proximal gradient step for minimizing $f(x) = g(x) + h(x)$ (page 4.4):

$$
\begin{aligned}
x_{k+1} &= \operatorname{prox}_{t_k h}(x_k - t_k \nabla g(x_k)) \\
&= \operatorname*{argmin}_u \left( h(u) + g(x_k) + \nabla g(x_k)^T (u - x_k) + \frac{1}{2t_k} \|u - x_k\|_2^2 \right)
\end{aligned}
$$

**Interpretation:** quadratic term represents

- a penalty that forces $x_{k+1}$ to be close to $x_k$, where linearization of $g$ is accurate
- an approximation of the error term in the linearization of $g$ at $x_k$

# Generalized proximal gradient method

replace $\frac{1}{2}\|u - x\|_2^2$ with a generalized distance $d(u, x)$:

$$x_{k+1} = \underset{u}{\operatorname{argmin}} \left( h(u) + g(x_k) + \nabla g(x_k)^T (u - x_k) + \frac{1}{t_k} d(u, x_k) \right)$$

## Potential benefits

1. "pre-conditioning": use a more accurate model of $g(u)$ around $x$, ideally

$$\frac{1}{t_k} d(u, x_k) \approx g(u) - g(x_k) - \nabla g(x_k)^T (u - x_k)$$

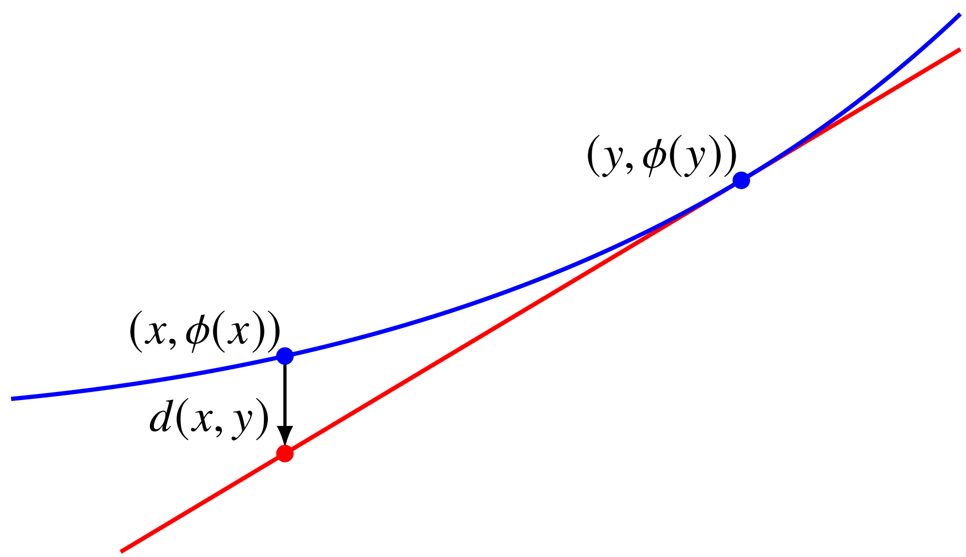
2. make the generalized proximal mapping (minimizer $u$) easier to compute

goal of 1 is to reduce number of iterations; goal of 2 is to reduce cost per iteration

# Bregman distance

$$d(x,y) = \phi(x) - \phi(y) - \nabla \phi(y)^T (x - y)$$

- $\phi$ is convex and continuously differentiable on $\operatorname{int}(\operatorname{dom}\phi)$
- domain of $\phi$ may include its boundary or a subset of its boundary
- we define the domain of $d$ as $\operatorname{dom} d = \operatorname{dom}\phi \times \operatorname{int}(\operatorname{dom}\phi)$
- $\phi$ is called the *kernel function* or *distance-generating function*




other properties of $\phi$ will be required but mentioned explicitly (*e.g.*, strict convexity)

# Immediate properties

$$d(x, y) = \phi(x) - \phi(y) - \nabla \phi(y)^T (x - y)$$

- $d(x, y)$ is convex in $x$ for fixed $y$
- $d(x, y) \geq 0$, with equality if $x = y$
- if $\phi$ is strictly convex, then $d(x, y) = 0$ only if $x = y$
- $d(x, y) \neq d(y, x)$ in general

to emphasize lack of symmetry, $d$ is also called a *directed distance* or *divergence*

# Examples

**Squared Euclidean distance** (with $\operatorname{dom}\phi = \mathbf{R}^n$)

$$\phi(x) = \frac{1}{2}x^T x, \qquad \nabla\phi(x) = x, \qquad d(x,y) = \frac{1}{2}\|x-y\|_2^2$$

**General quadratic kernel** (with $\operatorname{dom}\phi = \mathbf{R}^n$)

$$\phi(x) = \frac{1}{2}x^T Ax, \qquad \nabla\phi(x) = Ax, \qquad d(x,y) = \frac{1}{2}(x-y)^T A(x-y)$$

- $A$ is symmetric positive definite
- in some applications, $A$ is positive semidefinite, but not positive definite

# Examples

**Relative entropy** (with $\operatorname{dom}\phi = \mathbf{R}^n_+$)

$$\phi(x) = \sum_{i=1}^{n} x_i \log x_i, \qquad \nabla\phi(x) = \begin{bmatrix} \log x_1 + 1 \\ \vdots \\ \log x_n + 1 \end{bmatrix}$$

$$d(x, y) = \sum_{i=1}^{n} \left( x_i \log \frac{x_i}{y_i} - x_i + y_i \right)$$

**Logistic loss divergence** (with $\operatorname{dom}\phi = [0, 1]^n$)

$$\phi(x) = \sum_{i=1}^{n} \left( x_i \log x_i + (1 - x_i) \log(1 - x_i) \right), \qquad \nabla\phi(x) = \begin{bmatrix} \log(x_1/(1 - x_1)) \\ \vdots \\ \log(x_n/(1 - x_n)) \end{bmatrix}$$

$$d(x, y) = \sum_{i=1}^{n} \left( x_i \log \frac{x_i}{y_i} + (1 - x_i) \log \frac{1 - x_i}{1 - y_i} \right)$$

# Examples

**Hellinger divergence** (with $\operatorname{dom}\phi = [-1, 1]^n$)

$$\phi(x) = -\sum_{i=1}^{n} \sqrt{1 - x_i^2}, \qquad \nabla\phi(x) = \begin{bmatrix} x_1/\sqrt{1 - x_1^2} \\ \vdots \\ x_n/\sqrt{1 - x_n^2} \end{bmatrix}$$

$$d(x, y) = \sum_{i=1}^{n} \left( -\sqrt{1 - x_i^2} + \frac{1 - x_i y_i}{\sqrt{1 - y_i^2}} \right)$$

# Examples

**Logarithmic barrier** (with $\operatorname{dom}\phi = \mathbf{R}^n_{++}$)

$$\phi(x) = -\sum_{i=1}^n \log x_i, \qquad \nabla\phi(x) = \begin{bmatrix} -1/x_1 \\ \vdots \\ -1/x_n \end{bmatrix}, \qquad d(x,y) = \sum_{i=1}^n \left(\frac{x_i}{y_i} - \log\frac{x_i}{y_i} - 1\right)$$

$d(x,y)$ is sometimes called *Itakura–Saito* divergence

**Inverse barrier** (with $\operatorname{dom}\phi = \mathbf{R}^n_{++}$)

$$\phi(x) = \sum_{i=1}^n \frac{1}{x_i}, \qquad \nabla\phi(x) = \begin{bmatrix} -1/x_1^2 \\ \vdots \\ -1/x_n^2 \end{bmatrix}, \qquad d(x,y) = \sum_{i=1}^n \frac{1}{y_i}\left(\sqrt{\frac{x_i}{y_i}} - \sqrt{\frac{y_i}{x_i}}\right)^2$$

# Bregman distances for symmetric matrices

$$d(X,Y) = \phi(X) - \phi(Y) - \mathrm{tr}(\nabla\phi(Y)(X-Y))$$

- kernel $\phi$ is a convex function on $\mathbf{S}^n$, differentiable on $\mathrm{int}\,(\mathrm{dom}\,\phi)$
- domain of $d$ is $\mathrm{dom}\,d = \mathrm{dom}\,\phi \times \mathrm{int}\,(\mathrm{dom}\,\phi)$

**Relative entropy** (with $\mathrm{dom}\,\phi = \mathbf{S}^n_{++}$)

$$\phi(X) = -\log\det X, \qquad \nabla\phi(X) = -X^{-1}$$

$$d(X,Y) = \mathrm{tr}(XY^{-1}) - \log\det(XY^{-1}) - n$$

- $d(X,Y)$ is relative entropy between normal distributions $N(0,X)$ and $N(0,Y)$
- also known as *Kullback–Leibler divergence*

# Bregman distances for symmetric matrices

**Matrix entropy** (with $\mathrm{dom}\,\phi = \mathbf{S}^n_{++}$):

$$\phi(X) = \mathrm{tr}(X \log X), \qquad \nabla\phi(X) = I + \log X$$

$$d(X,Y) = \mathrm{tr}(X \log X - X \log Y - X + Y)$$

- matrix logarithm $\log X$ is defined as

$$\log X = \sum_{i=1}^{n} (\log \lambda_i) q_i q_i^T$$

  if $X$ has eigendecomposition $X = \sum_i \lambda_i q_i q_i^T$

- $d(X,Y)$ is also known as *quantum relative entropy*

# Outline

- Bregman distance
- **properties**
- Bregman proximal mapping
- mirror descent

# Three-point identity

for all $x \in \operatorname{dom}\phi$ and $y, z \in \operatorname{int}(\operatorname{dom}\phi)$,

$$d(x,z) = d(x,y) + d(y,z) + (\nabla\phi(y) - \nabla\phi(z))^T(x-y)$$

- easily verified by substituting the definition of $d$
- if $d$ is not symmetric, order of the arguments of $d$ in the identity matters
- generalizes the familiar identity for squared Euclidean distance:

$$\frac{1}{2}\|x-z\|_2^2 = \frac{1}{2}\|x-y\|_2^2 + \frac{1}{2}\|y-z\|_2^2 + (y-z)^T(x-y)$$

# Strongly convex kernel

we will sometimes assume that $\phi$ is strongly convex (page 1.19):

$$\phi(x) \geq \phi(y) + \nabla\phi(y)^T(x-y) + \frac{\mu}{2}\|x-y\|^2$$

- $\mu > 0$ is strong convexity constant of $\phi$ for the norm $\|\cdot\|$
- for twice differentiable $\phi$, this is equivalent to

$$v^T\nabla^2\phi(x)v \geq \mu\|v\|^2 \quad \text{for all } x \in \operatorname{int}(\operatorname{dom}\phi) \text{ and } v$$

  (see page 1.18)
- strong convexity of $\phi$ implies that

$$\begin{aligned} d(x,y) &= \phi(x) - \phi(y) - \nabla\phi(y)^T(x-y) \\ &\geq \frac{\mu}{2}\|x-y\|^2 \end{aligned}$$

# Regularization with Bregman distance

for given $y \in \mathrm{int}(\mathrm{dom}\, \phi)$ and convex $f$, consider

$$\text{minimize} \quad f(x) + d(x, y)$$

- equivalently, minimize $f(x) + \phi(x) - \nabla \phi(y)^T x$
- feasible set is $\mathrm{dom}\, f \cap \mathrm{dom}\, \phi$

**Optimality condition:** $\hat{x} \in \mathrm{dom}\, f \cap \mathrm{int}(\mathrm{dom}\, \phi)$ is optimal if and only if

$$f(x) + d(x, y) \geq f(\hat{x}) + d(\hat{x}, y) + d(x, \hat{x}) \quad \text{for all } x \in \mathrm{dom}\, f \cap \mathrm{dom}\, \phi \qquad (1)$$

**Equivalent optimality condition:** $\hat{x} \in \mathrm{dom}\, f \cap \mathrm{int}(\mathrm{dom}\, \phi)$ is optimal if and only if

$$\nabla \phi(y) - \nabla \phi(\hat{x}) \in \partial f(\hat{x}) \qquad (2)$$

*Proof:* we derive optimality conditions for the problem

$$\text{minimize} \quad g(x) + \phi(x) \tag{3}$$

with $g$ convex, and apply the results to $g(x) = f(x) - \nabla\phi(y)^T x$

- optimality condition: $\hat{x} \in \operatorname{dom} g \cap \operatorname{int}(\operatorname{dom} \phi)$ is optimal for (3) if and only if

$$g(x) \geq g(\hat{x}) - \nabla\phi(\hat{x})^T(x - \hat{x}) \quad \text{for all } x \in \operatorname{dom} g \cap \operatorname{dom} \phi \tag{4}$$

  combined with the 3-point identity this gives the optimality condition (1)

- equivalent optimality condition: $\hat{x} \in \operatorname{dom} g \cap \operatorname{int}(\operatorname{dom} \phi)$ is optimal if and only if

$$-\nabla\phi(\hat{x}) \in \partial g(\hat{x}) \tag{5}$$

  applied to $g(x) = f(x) - \nabla\phi(y)^T x$ this gives the optimality condition (2)

*Proof:*

$$\begin{tikzcd}[column sep=small] & \text{optimality of } \hat{x} \arrow[dr, Rightarrow, "c"] & \\ (4) \arrow[ur, Rightarrow, "a"] & & (5) \arrow[ll, Rightarrow, "b"'] \end{tikzcd}$$

- implication a follows from convexity of $\phi$: if (4) holds, then for all feasible $x$,

$$g(x) + \phi(x) \geq g(\hat{x}) + \phi(x) - \nabla\phi(\hat{x})^T(x - \hat{x}) \geq g(\hat{x}) + \phi(\hat{x})$$

- implication b: by definition of subgradient, (5) can be written as

$$g(x) \geq g(\hat{x}) - \nabla\phi(\hat{x})^T(x - \hat{x}) \quad \text{for all } x \in \operatorname{dom} g$$

- we prove c by contradiction: suppose that for some $x \in \operatorname{dom} g$

$$g(x) < g(\hat{x}) - \nabla\phi(\hat{x})^T(x - \hat{x})$$

define $v = x - \hat{x}$; for small positive $t$, by convexity of $g$ and Taylor's theorem,

$$\begin{aligned} g(\hat{x} + tv) + \phi(\hat{x} + tv) &\leq g(\hat{x}) + t(g(x) - g(\hat{x})) + \phi(\hat{x} + tv) \\ &= g(\hat{x}) + \phi(\hat{x}) + t(g(x) - g(\hat{x}) + \nabla\phi(\hat{x})^T v) + O(t^2) \\ &< g(\hat{x}) + \phi(\hat{x}) \end{aligned}$$

# Outline

- Bregman distance
- properties
- **Bregman proximal mapping**
- mirror descent

# Bregman proximal mapping

for convex $f$ and Bregman kernel $\phi$, define

$$\begin{aligned} \operatorname{prox}_f^d(y,a) &= \operatorname*{argmin}_x \left( f(x) + a^T x + d(x,y) \right) \\ &= \operatorname*{argmin}_x \left( f(x) + (a - \nabla\phi(y))^T x + \phi(x) \right) \end{aligned}$$

- first argument $y$ must be in $\operatorname{int}(\operatorname{dom}\phi)$
- second argument $a$ can take any value
- we'll use this only if for every $y$ and $a$, a unique minimizer $x \in \operatorname{int}(\operatorname{dom}\phi)$ exists

# Example: quadratic kernel

$$\phi(x) = \frac{1}{2}\|x\|_2^2, \qquad d(x,y) = \frac{1}{2}\|x-y\|_2^2$$

Bregman proximal mapping can be expressed in terms of standard $\mathrm{prox}_f$:

$$\begin{aligned}
\mathrm{prox}_f^d(y,a) &= \underset{x}{\mathrm{argmin}} \left( f(x) + a^T x + d(x,y) \right) \\
&= \underset{x}{\mathrm{argmin}} \left( f(x) + a^T x + \frac{1}{2}\|x-y\|_2^2 \right) \\
&= \mathrm{prox}_f(y-a)
\end{aligned}$$

closedness of $f$ ensures existence and uniqueness (see page 6.2)

# Example: relative entropy

$$\phi(x) = \sum_{i=1}^{n} x_i \log x_i, \qquad d(x,y) = \sum_{i=1}^{n} (x_i \log(x_i/y_i) - x_i + y_i)$$

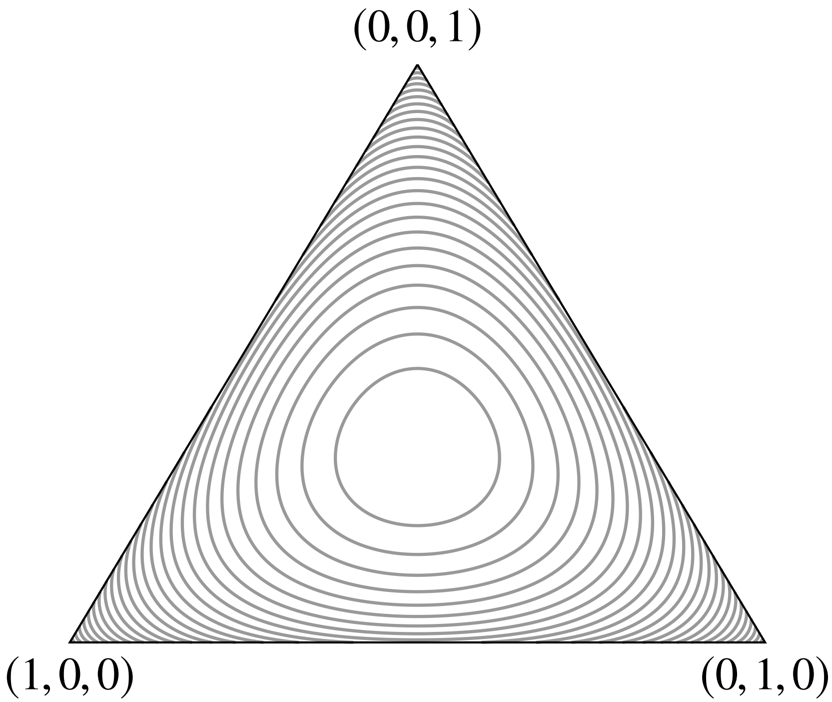
- we take $f = \delta_C$, the indicator of probability simplex $C = \{x \succeq 0 \mid \mathbf{1}^T x = 1\}$
- Bregman proximal mapping is

$$\begin{aligned}
\mathrm{prox}_f^d(y,a) &= \operatorname*{argmin}_{\mathbf{1}^T x = 1} \left(a^T x + \sum_{i=1}^{n} x_i \log(x_i/y_i)\right) \\
&= \frac{1}{\sum_{i=1}^{n} y_i e^{-a_i}} \begin{bmatrix} y_1 e^{-a_1} \\ \vdots \\ y_n e^{-a_n} \end{bmatrix}
\end{aligned}$$

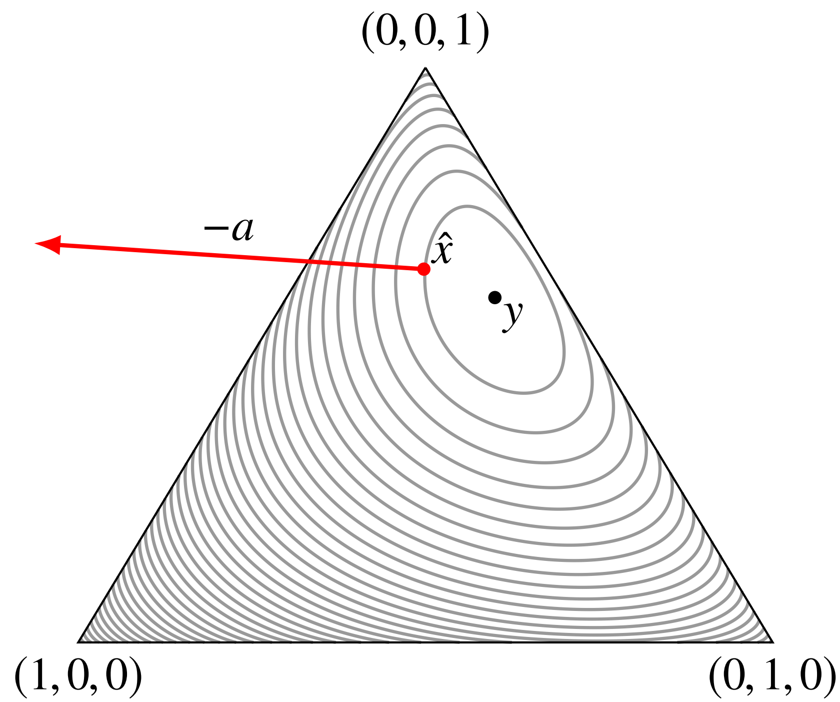
- for every $y \succ 0$ and $a$, minimizer in the definition exists, is unique, and positive

# Example: relative entropy

Contour lines of $\phi(x)$




Contour lines of $d(x, y)$




right-hand figure shows

$$\hat{x} = \mathrm{prox}_f^d(y, a) = \mathrm{argmin}\,(a^T x + d(x, y))$$

for $y = (0.1,\ 0.3,\ 0.6)$ and $a = (-0.540,\ 0.585,\ -0.045)$

# Optimality condition

apply the optimality conditions for Bregman-regularized problem (page 13.14) to

$$\mathrm{prox}_f^d(y, a) = \underset{x}{\mathrm{argmin}} \left( f(x) + a^T x + d(x, y) \right)$$

suppose $\hat{x} \in \mathrm{dom}\, f \cap \mathrm{int}\,(\mathrm{dom}\, \phi)$

- first condition: $\hat{x} = \mathrm{prox}_f^d(y, a)$ if and only if

$$f(x) + a^T x + d(x, y) \geq f(\hat{x}) + a^T \hat{x} + d(\hat{x}, y) + d(x, \hat{x})$$

  for all $x \in \mathrm{dom}\, f \cap \mathrm{dom}\, \phi$

- second condition: $\hat{x} = \mathrm{prox}_f^d(y, a)$ if and only if

$$\nabla \phi(y) - \nabla \phi(\hat{x}) - a \in \partial f(\hat{x})$$

# Outline

- Bregman distance
- properties
- Bregman proximal mapping
- **mirror descent**

# Mirror descent

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in C \end{array}$$

- $f$ is a convex function, $C$ is a convex subset of $\mathrm{dom}\, f$
- we assume $f$ is subdifferentiable on $C$

**Algorithm:** choose $x_0 \in C \cap \mathrm{int}(\mathrm{dom}\, \phi)$ and repeat

$$x_{k+1} = \operatorname*{argmin}_{x \in C} \left( t_k g_k^T x + d(x, x_k) \right), \quad k = 0, 1, \ldots$$

$g_k$ is any subgradient of $f$ at $x_k$

update can be written as $x_{k+1} = \mathrm{prox}_{\delta_C}^{d}(x_k, t_k g_k)$ where $\delta_C$ is indicator of $C$

# Mirror descent with quadratic kernel

$$x_{k+1} = \underset{x \in C}{\operatorname{argmin}} \left( t_k g_k^T x + d(x, x_k) \right)$$

for $d(x,y) = \frac{1}{2}\|x - y\|_2^2$, this is the projected subgradient method:

$$\begin{aligned}
x_{k+1} &= \underset{x \in C}{\operatorname{argmin}} \left( t g_k^T x + \frac{1}{2}\|x - x_k\|_2^2 \right) \\
&= \underset{x \in C}{\operatorname{argmin}} \frac{1}{2}\|x - x_k + t_k g_k\|_2^2 \\
&= P_C(x_k - t_k g_k)
\end{aligned}$$

where $P_C$ is Euclidean projection on $C$

# Assumptions

- problem on page 13.22 has optimal value $f^\star$, optimal solution $x^\star \in C \cap \operatorname{dom} \phi$
- $f$ is Lipschitz continuous on $C$ with respect to some norm $\|\cdot\|$

$$|f(x) - f(y)| \leq G\|x - y\| \quad \text{for all } x, y \in C$$

  this is equivalent to $\|g\|_* \leq G$ for all $x \in C$ and $g \in \partial f(x)$

  (proof extends proof for Euclidean norm on page 3.4)

- $\phi$ is $1$-strongly convex on $C$, with respect to the same norm $\|\cdot\|$:

$$d(x, y) \geq \frac{1}{2}\|x - y\|^2 \quad \text{for all } x \in \operatorname{dom} \phi \cap C \text{ and } y \in \operatorname{int}(\operatorname{dom} \phi) \cap C$$

# Analysis

- apply optimality condition on page 13.21 with $x = x^\star$, $y = x_i$, $\hat{x} = x_{i+1}$:

$$\begin{aligned}
d(x^\star, x_{i+1}) &\leq d(x^\star, x_i) - d(x_{i+1}, x_i) + t_i g_i^T (x_i - x_{i+1}) + t_i g_i^T (x^\star - x_i) \\
&\leq d(x^\star, x_i) - d(x_{i+1}, x_i) + \|t_i g_i\|_* \|x_{i+1} - x_i\| + t_i g_i^T (x^\star - x_i) \\
&\leq d(x^\star, x_i) - d(x_{i+1}, x_i) + \frac{1}{2}\|x_{i+1} - x_i\|^2 + \frac{1}{2}\|t_i g_i\|_*^2 + t_i g_i^T (x^\star - x_i)
\end{aligned}$$

  last step is arithmetic-geometric mean inequality

- apply strong convexity of kernel and definition of subgradient:

$$d(x^\star, x_{i+1}) \leq d(x^\star, x_i) + \frac{1}{2}\|t_i g_i\|_*^2 + t_i(f^\star - f(x_i))$$

- define $f_{\text{best},k} = \min_{i=0,\ldots,k} f(x_i)$ and combine inequalities for $i = 0, \ldots, k$:

$$\begin{aligned}
\left(\sum_{i=0}^{k} t_i\right)(f_{\text{best},k} - f^\star) &\leq d(x^\star, x_0) - d(x^\star, x_{k+1}) + \frac{1}{2}\sum_{i=0}^{k} \|t_i g_i\|_*^2 \\
&\leq d(x^\star, x_0) + \frac{1}{2}\sum_{i=0}^{k} \|t_i g_i\|_*^2
\end{aligned}$$

# Step size selection

$$f_{\mathrm{best},k} - f^\star \le \frac{d(x^\star, x_0)}{\sum_{i=0}^{k} t_i} + \frac{\sum_{i=0}^{k} \|t_i g_i\|_*^2}{2\sum_{i=0}^{k} t_i} \le \frac{d(x^\star, x_0)}{\sum_{i=0}^{k} t_i} + \frac{G^2 \sum_{i=0}^{k} t_i^2}{2\sum_{i=0}^{k} t_i}$$

- diminishing step size: $f_{\mathrm{best},k} \to f^\star$ if

$$t_i \to 0, \qquad \sum_{i=0}^{\infty} t_i = \infty$$

(see page 3.7)

- optimal step size for fixed number of iterations $k$, if we know that $d(x^\star, x_0) \le D$:

$$t_i = \frac{\sqrt{2D}}{\|g_i\|_* \sqrt{k+1}}, \qquad f_{\mathrm{best},k} \le \frac{G\sqrt{2D}}{\sqrt{k+1}}$$

(see page 3.10)

# Entropic mirror descent

apply mirror descent with relative entropy distance and

$$C = \{x \in \mathbf{R}^n \mid x \succeq 0,\ \mathbf{1}^T x = 1\}$$

**Algorithm:** choose $x_0 \succ 0$, $\mathbf{1}^T x_0 = 1$, and repeat

$$x_{k+1} = \frac{1}{s^T x_k}(s \circ x_k) \quad \text{where } s = \left(e^{-t_k g_{k,1}}, \ldots, e^{-t_k g_{k,n}}\right)$$

- $g_k$ is any subgradient of $f$ at $x_k$
- $\circ$ denotes component-wise vector product

# Convergence

in the analysis on page 13.26

- if we choose $x_0 = (1/n)\mathbf{1}$, then we can take $D = \log n$:

$$d(x^\star, x_0) = \log n + \sum_{i=1}^{n} x_i^\star \log x_i^\star \leq \log n$$

- $\phi(x) = \sum\limits_i x_i \log x_i$ is 1-strongly convex for $\|\cdot\|_1$ on $C$: by Cauchy–Schwarz,

$$v^T \nabla^2 \phi(x) v = \sum_{i=1}^{n} \frac{v_i^2}{x_i} \geq \|v\|_1^2 \qquad \text{if } x \succ 0, \quad \mathbf{1}^T x = 1$$

- with optimal step size for $k$ iterations,

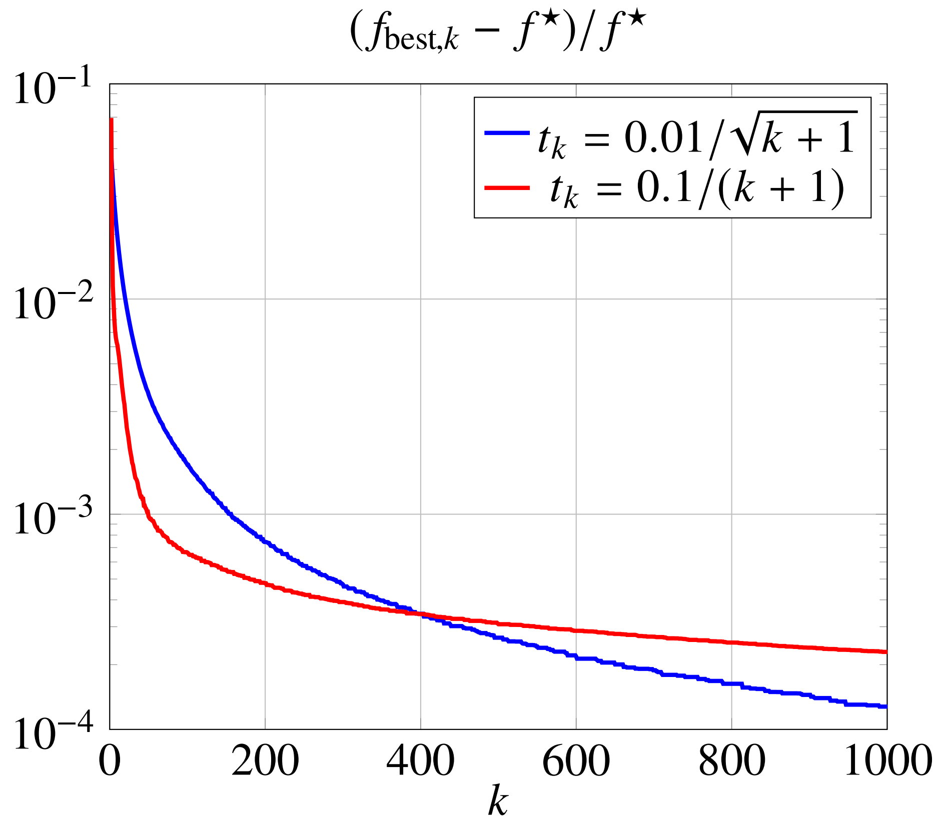
$$f_{\mathrm{best},k} \leq \frac{G\sqrt{2\log n}}{\sqrt{k+1}}$$

where $G$ is Lipschitz constant of $f$ for $\|\cdot\|_1$-norm

# Example

$$\begin{array}{ll} \text{minimize} & \|Ax - b\|_1 \\ \text{subject to} & x \succeq 0, \quad \mathbf{1}^T x = 1 \end{array}$$

- subgradient $g = A^T \text{sign}(Ax - b)$, so $\|g\|_\infty \leq G = \max_j \sum_i |A_{ij}|$
- example with randomly generated $A \in \mathbf{R}^{1000 \times 500}$, $b \in \mathbf{R}^{1000}$



# References

## Generalized distances


- Y. Censor and S. A. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications* (1997).
- M. Basseville, *Distance measures for statistical data processing—An annotated bibliography*, Signal Processing (2013).


## Mirror descent


- A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization* (1983).
- A. Beck and M. Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Operations Research Letters (2003).
- A. Juditsky and A. Nemirovski, *First-order methods for nonsmooth convex large-scale optimization, I: General-purpose methods*. In S. Sra, S. Nowozin, S. J. Wright, editors, *Optimization for Machine Learning* (2012).
- A. Beck, *First-Order Methods in Optimization* (2017), chapter 9.