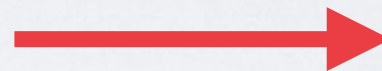


PROXIMAL METHODS

WHY PROXIMAL METHODS

Smooth functions
minimize $f(x)$



Gradient descent
Newton's method
Quasi-newton
Conjugate gradients
etc...

Non-differentiable
minimize $f(x)$



Proximal methods

Constrained problems?

minimize $f(x)$
subject to $g(x) \leq 0$
 $h(x) = 0$



Lagrangian methods

PROXIMAL OPERATOR

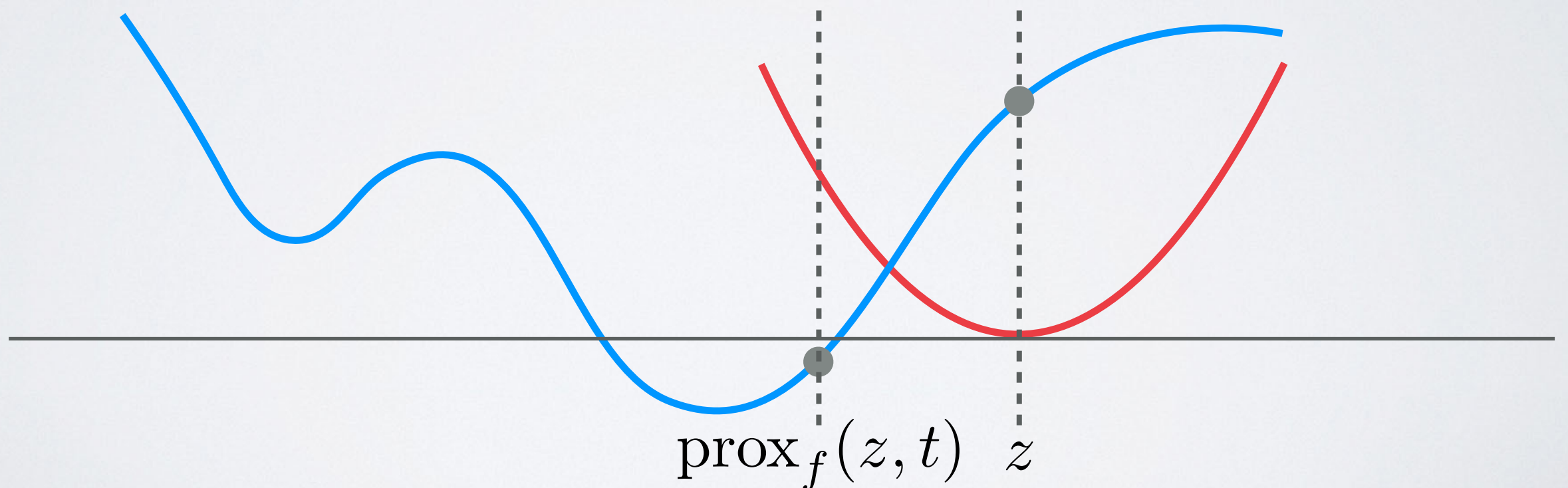
$$\text{prox}_f(z, \tau) = \arg \min_x f(x) + \underbrace{\frac{1}{2\tau} \|x - z\|^2}_{\substack{\text{proximal} \\ \text{penalty} \\ \text{"Stay close to } z\text{"}}}$$

objective 

PROXIMAL OPERATOR

$$\text{prox}_f(z, \tau) = \arg \min_x f(x) + \frac{1}{2\tau} \|x - z\|^2$$

stepsize



GRADIENT INTERPRETATION

$$\text{prox}_f(z, \tau) = \arg \min_x f(x) + \frac{1}{2\tau} \|x - z\|^2$$

assume differentiable

$$\nabla f(x) + \frac{1}{\tau}(x - z) = 0$$

$$x = z - \tau \nabla f(x)$$

Two types of gradient descent

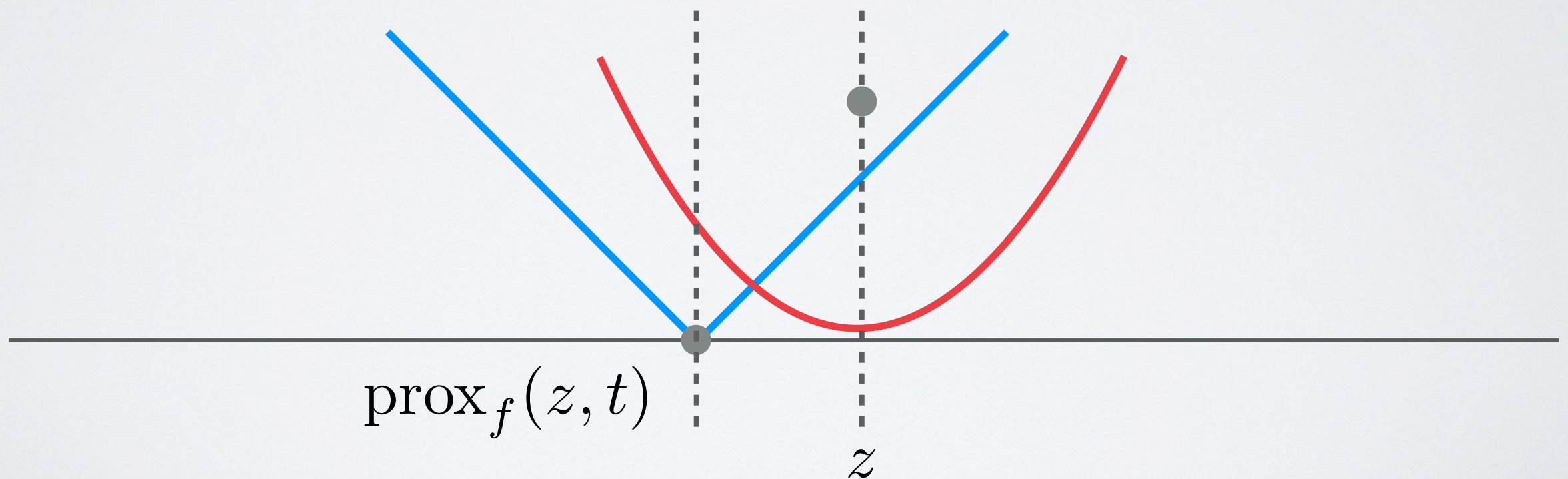
Forward gradient $x = z - \tau \nabla f(z)$

Backward gradient $x = z - \tau \nabla f(x)$

EXAMPLE: L1

$$\text{prox}_f(z, \tau) = \arg \min_x |x| + \frac{1}{2\tau} \|x - z\|^2$$

Is sub-differential unique? Is solution unique?



PROPERTIES

backward gradient descent

forward gradient descent

$$\text{prox}_f(z, \tau) = \arg \min_x f(x) + \frac{1}{2\tau} \|x - z\|^2$$

$$x = z - \tau \partial f(x)$$

$$x = z - \tau \nabla f(z)$$

Existence

Exists for any proper
convex function
(and some non-convex)

Exists only for smooth
functions

Uniqueness

Always unique result

Sub-gradient unique if
differentiable

EXAMPLE: L1

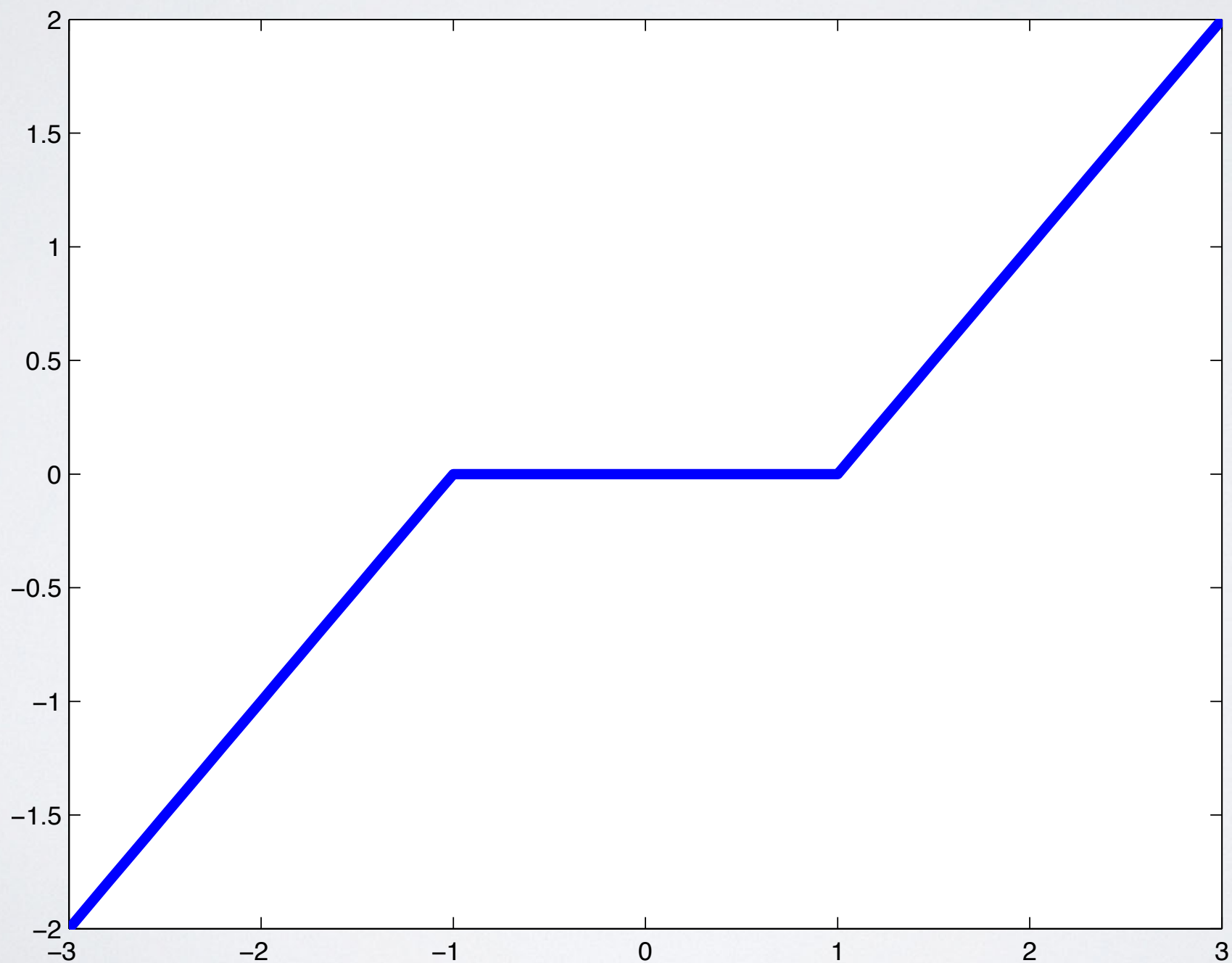
$$\text{prox}_f(z, \tau) = \arg \min_x |x| + \frac{1}{2\tau} \|x - z\|^2$$

$$\begin{cases} 1 + \frac{1}{\tau}(x^* - z) = 0, & \text{if } x^* > 0 \\ -1 + \frac{1}{\tau}(x^* - z) = 0, & \text{if } x^* < 0 \\ \frac{1}{\tau}(x^* - z) \in [-1, 1], & \text{if } x^* = 0 \end{cases}$$

$$\text{shrink}(z, \tau) = \begin{cases} x^* = z - \tau, & \text{if } x^* > 0 \\ x^* = z + \tau, & \text{if } x^* < 0 \\ 0, & \text{otherwise} \end{cases}$$

SHRINK OPERATOR

$$y = \text{shrink}(x, 1)$$



NUCLEAR NORM

$$\|X\|_* = \sum |\lambda_i|$$

Singular values



Why would you use this regularizer?

If X is PSD, we call this the “**trace norm**”

$$\|X\|_* = \sum \lambda_i = \text{trace}(X)$$

NUCLEAR NORM PROX

$$\text{minimize} \quad \|X\|_* + \frac{1}{2\tau} \|X - Z\|^2$$

Same singular vectors

$$\text{minimize} \quad \|US_X V^T\|_* + \frac{1}{2\tau} \|US_X V^T - US_Z V^T\|^2$$

$$\text{minimize} \quad |S_X| + \frac{1}{2} \|S_X - S_Z\|^2$$

$$S_X = \text{shrink}(S_Z, \tau)$$

$$X = U \text{shrink}(S_Z, \tau) V^T$$

CHARACTERISTIC FUNCTIONS

C = some convex set

$$\chi_C(x) = \begin{cases} 0, & \text{if } x \in C \\ \infty, & \text{otherwise} \end{cases}$$

proximal

$$\text{minimize} \quad \chi_C(x) + \frac{1}{2t} \|x - z\|^2$$

What does this do?

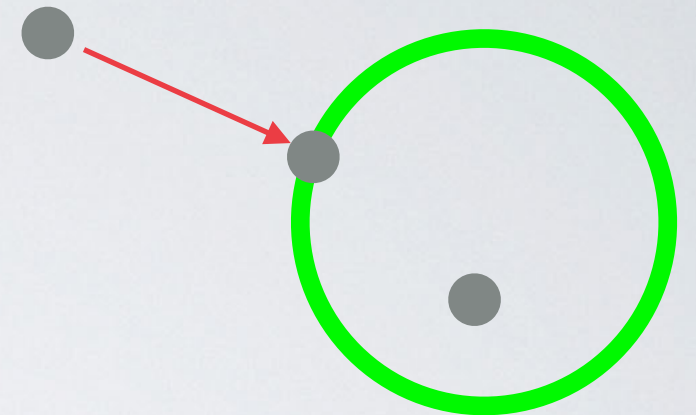
Does the stepsize matter?

EXAMPLES

2-norm ball

$$C = B_2 = \{x \mid \|x\| \leq 1\}$$

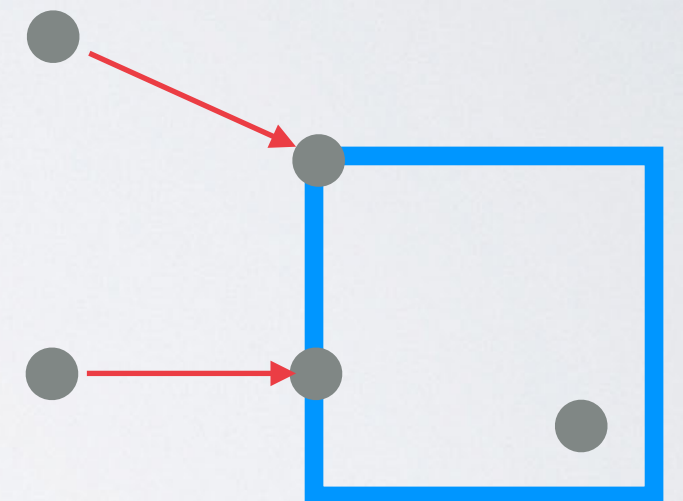
$$\text{prox}_2(z) = \frac{z}{\|z\|} \min\{\|z\|, 1\}$$



infinity ball

$$C = B_\infty = \{x \mid \|x\|_\infty \leq 1\}$$

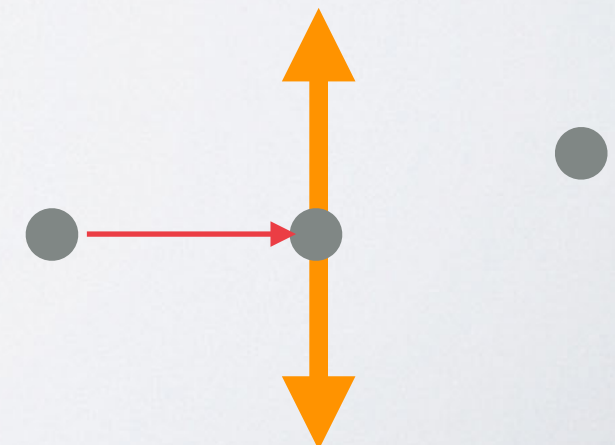
$$\text{prox}_\infty(z)_i = \min\{\max\{z_i, -1\}, 1\}$$



positive half space

$$C = \{x \mid x \geq 0\}$$

$$\text{prox}_+(z) = \max\{z, 0\}$$

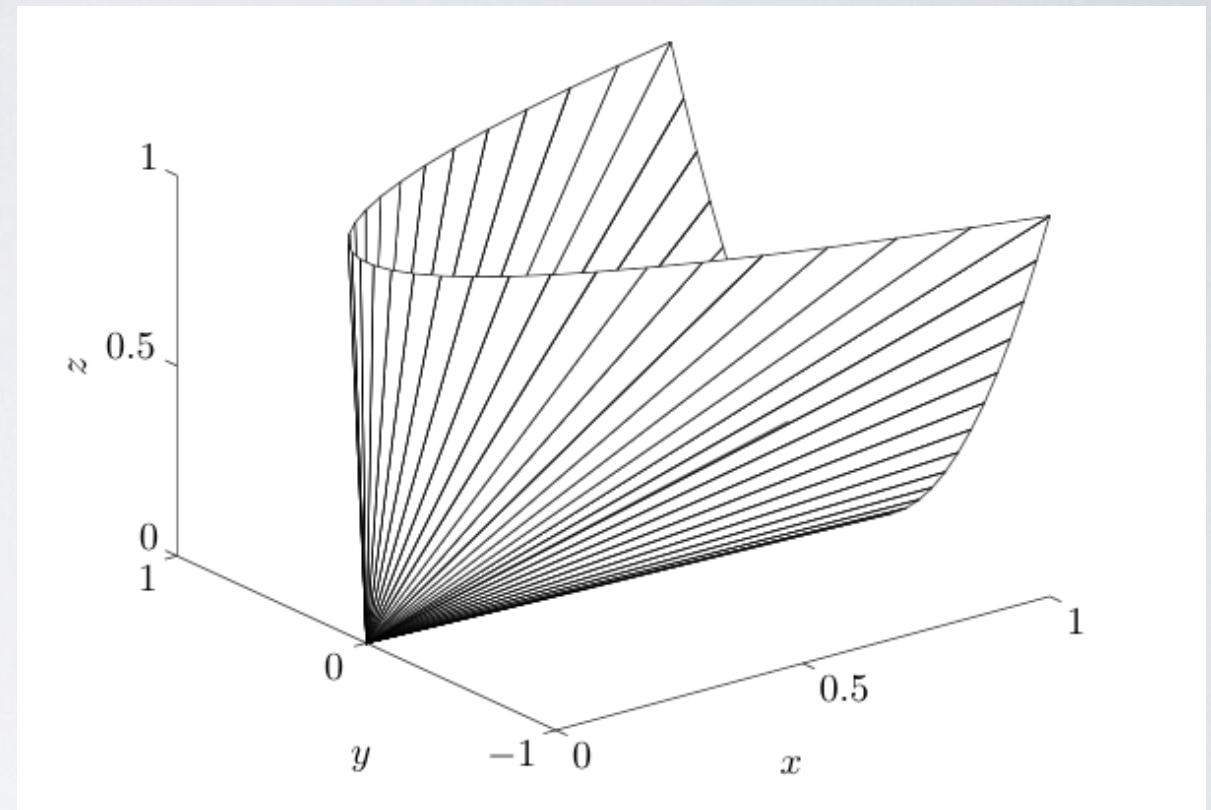


HARDER EXAMPLES

semidefinite cone

$$C = S_{++} = \{X | X \succeq 0\}$$

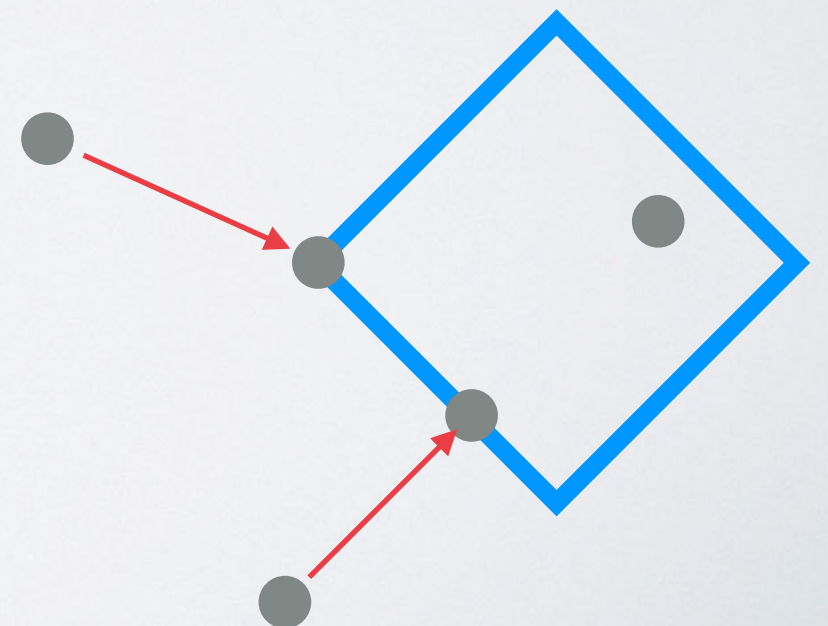
$$\text{prox}_{S_{++}}(Z) = U \max\{S, 0\} V^T$$



L1-ball

$$C = B_1 = \{x \mid \|x\|_1 \leq 1\}$$

$$\text{prox}_1(z) = ???$$



PROXIMAL-POINT METHOD

$$\text{minimize } f(x)$$

backward gradient descent

$$x^{k+1} = \text{prox}_f(x^k, \tau) = x^k - \tau \partial f(x^{k+1})$$

$$x^{k+1} = \arg \min f(x) + \frac{1}{2\tau} \|x - x^k\|^2$$

stepsize restriction?

does it depend on the norm?

Would you ever use this?

FORWARD-BACKWARD SPLITTING

$$\underset{x}{\text{minimize}} \quad g(x) + f(x)$$

non-smooth

smooth

backward gradient descent

forward gradient descent

FBS

forward step

$$\hat{x} = x^k - \tau \nabla f(x^k)$$

backward step

$$x^{k+1} = \text{prox}_g(\hat{x}, \tau)$$

WHY FORWARD BACKWARD?

$$\underset{x}{\text{minimize}} \quad g(x) + f(x)$$

forward step $\hat{x} = x^k - \tau \nabla f(x^k)$

backward step $x^{k+1} = \text{prox}_g(\hat{x}, \tau)$

$$x^{k+1} = x^k - \tau \nabla f(x^k) - \tau \partial g(x^{k+1})$$

fixed-point property

$$x^* = x^* - \tau \nabla f(x^*) - \tau \partial g(x^*)$$

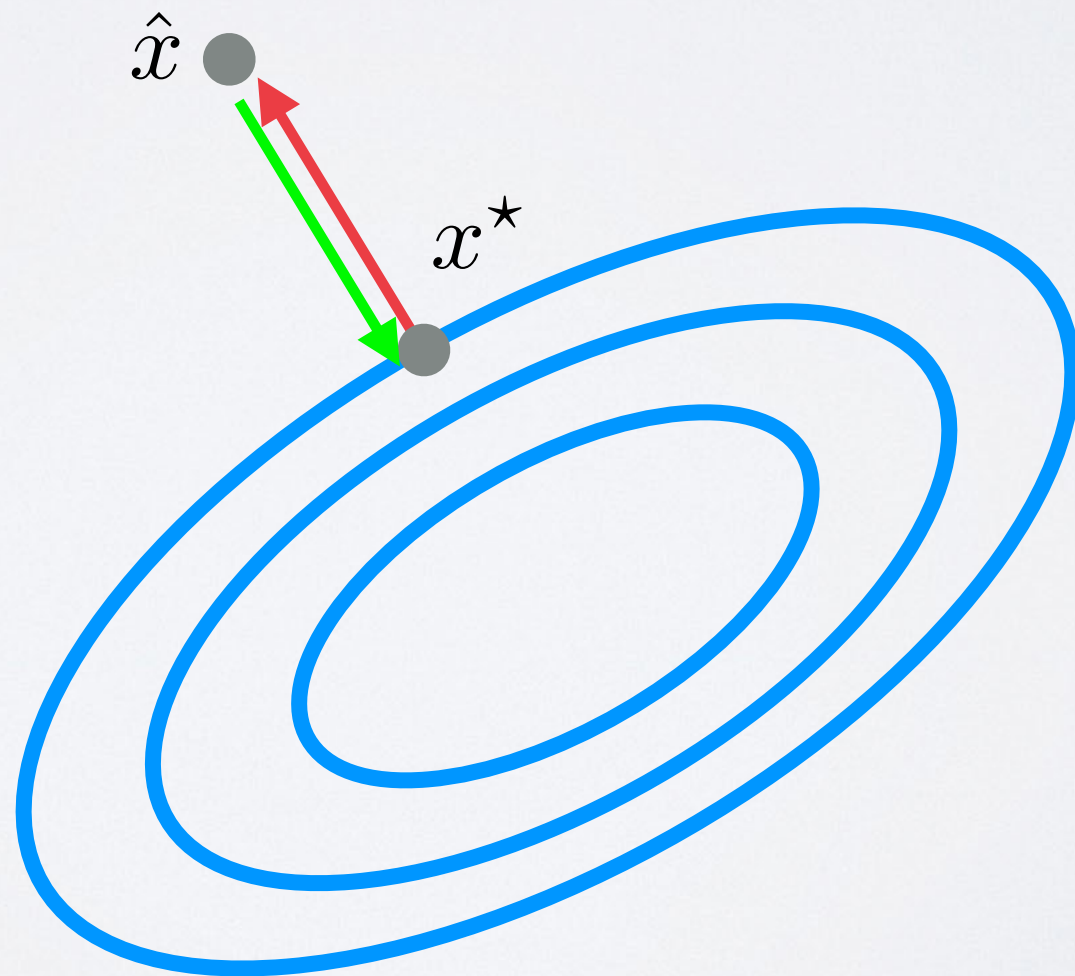
$$0 \in \nabla f(x^*) + \partial g(x^*)$$

GRADIENT INTERPRETATION

forward step $\hat{x} = x^k - \tau \nabla f(x^k)$

backward step $x^{k+1} = \text{prox}_g(\hat{x}, \tau)$

$$x^{k+1} = x^k - \nabla f(x^k) - \partial g(x^{k+1})$$



MAJORIZATION-MINIMIZATION

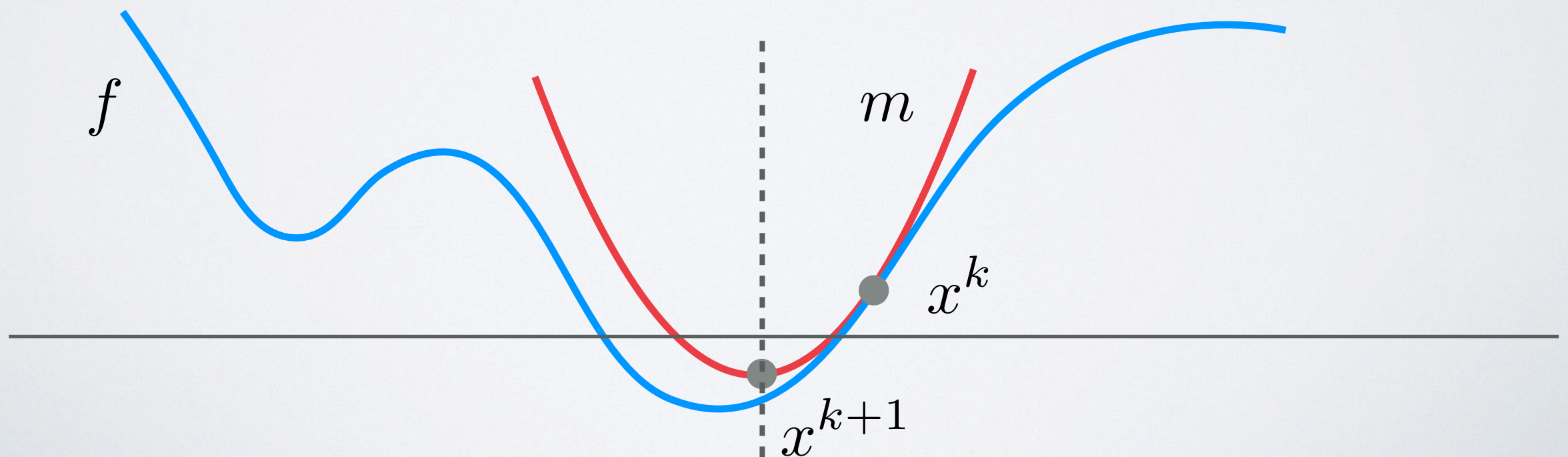
$$\text{minimize } h(x)$$

surrogate function

$$m(x^k) = h(x^k)$$

$$m(x) \geq h(x), \quad \forall x$$

surrogate “majorizes” f

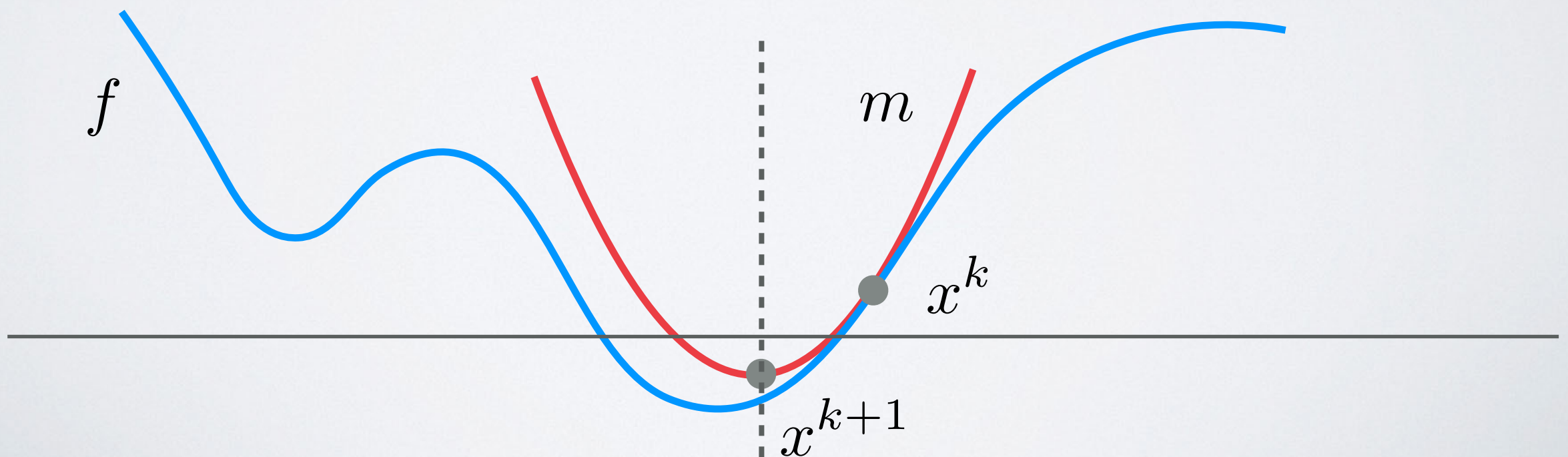


MM ALGORITHM

$$\text{minimize} \quad h(x) = g(x) + f(x)$$

$$f(x) \leq f(x^k) + \langle x - x^k, \nabla f(x^k) \rangle + \frac{1}{2\tau} \|x - x^k\|^2$$


$$\tau \leq \frac{1}{L}$$



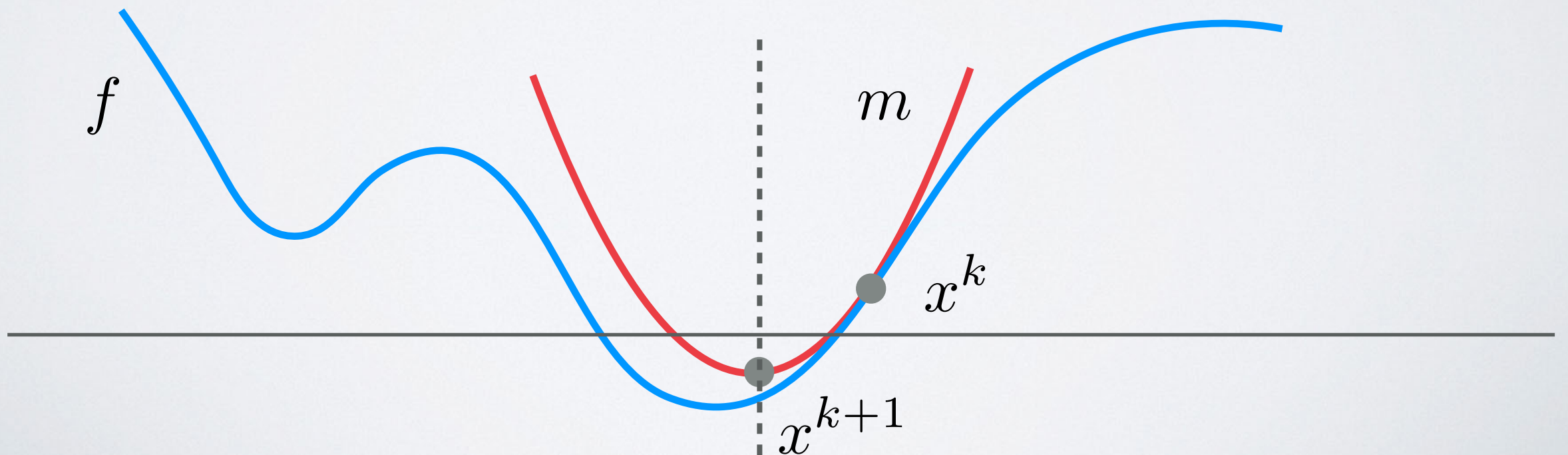
MM ALGORITHM

$$\text{minimize} \quad h(x) = g(x) + f(x)$$

$$f(x) \leq f(x^k) + \langle x - x^k, \nabla f(x^k) \rangle + \frac{1}{2\tau} \|x - x^k\|^2$$

$$\text{minimize} \quad m(x) = g(x) + f(x^k) + \langle x - x^k, \nabla f(x^k) \rangle + \frac{1}{2\tau} \|x - x^k\|^2$$

$$\text{minimize} \quad m(x) = g(x) + \frac{1}{2\tau} \|x - x^k + \tau \nabla f(x^k)\|^2$$



MM ALGORITHM

$$\text{minimize} \quad h(x) = g(x) + f(x)$$

$$f(x) \leq f(x^k) + \langle x - x^k, \nabla f(x^k) \rangle + \frac{1}{2\tau} \|x - x^k\|^2$$

$$\text{minimize} \quad m(x) = g(x) + f(x^k) + \langle x - x^k, \nabla f(x^k) \rangle + \frac{1}{2\tau} \|x - x^k\|^2$$

$$\text{minimize} \quad m(x) = g(x) + \frac{1}{2\tau} \|x - x^k + \tau \nabla f(x^k)\|^2$$

$$\text{prox}(x^k - \tau \nabla f(x^k), \tau)$$

SPARSE LEAST SQUARES

$$\text{minimize } g(x) + f(x)$$

$$\text{minimize } \mu|x| + \frac{1}{2}\|Ax - b\|^2$$

non-smooth

smooth

forward step

$$\hat{x} = x^k - \tau \nabla f(x^k)$$

$$\hat{x} = x^k - \tau A^T (Ax^k - b)$$

backward step

$$x^{k+1} = \arg \min g(x) + \frac{1}{2\tau} \|x - \hat{x}\|^2$$

$$x^{k+1} = \arg \min \mu|x| + \frac{1}{2\tau} \|x - \hat{x}\|^2$$

$$x^{k+1} = \text{shrink}(\hat{x}, \mu\tau)$$

iterative shrinkage / thresholding

SPARSE LOGISTIC

$$\text{minimize } g(x) + f(x)$$

$$\text{minimize } \mu|x| + L(Ax, y)$$

$$L(z, y) = \sum_i \log(1 + e^{-y_i z_i})$$

forward step

$$\hat{x} = x^k - \tau \nabla f(x^k)$$

$$\hat{x} = x^k - \tau A^T \nabla L(Ax^k, y)$$

backward step

$$x^{k+1} = \arg \min g(x) + \frac{1}{2\tau} \|x - \hat{x}\|^2$$

$$x^{k+1} = \arg \min \mu|x| + \frac{1}{2\tau} \|x - \hat{x}\|^2$$


$$x^{k+1} = \text{shrink}(\hat{x}, \mu\tau)$$

EXAMPLE: DEMOCRATIC REPRESENTATIONS

$$\left\{ \begin{array}{l} \text{minimize} \quad \frac{1}{2} \|Ax - b\|^2 \\ \text{subject to} \quad \|x\|_\infty \leq \mu \end{array} \right.$$

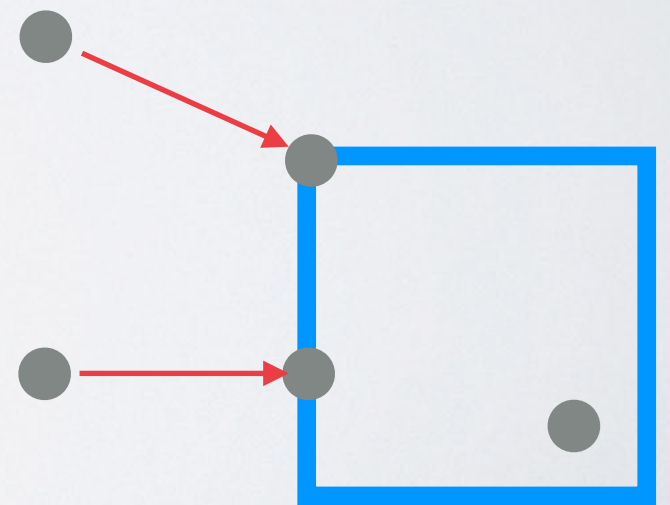
minimize $\mathcal{X}_\infty^\mu(x) + \frac{1}{2} \|Ax - b\|^2$

forward step

$$\hat{x} = x^k - \tau A^T (Ax^k - b)$$

backward step

$$\begin{aligned} x^{k+1} &= \arg \min \mathcal{X}_\infty^\mu(x) + \frac{1}{2\tau} \|x - \hat{x}\|^2 \\ &= \min\{\max\{\hat{x}, -\mu\}, \mu\} \end{aligned}$$



EXAMPLE: LASSO

$$\left\{ \begin{array}{l} \text{minimize} \quad \frac{1}{2} \|Ax - b\|^2 \\ \text{subject to} \quad |x| \leq \mu \end{array} \right.$$

implicit constraints

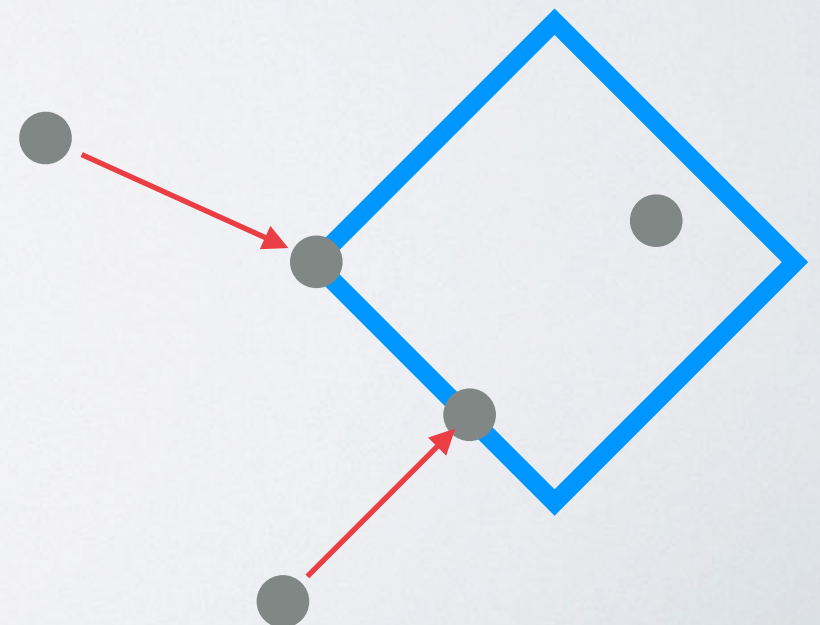
$$\text{minimize} \quad \mathcal{X}_1^\mu(x) + \frac{1}{2} \|Ax - b\|^2$$

forward step

$$\hat{x} = x^k - \tau A^T (Ax^k - b)$$

backward step

$$x^{k+1} = \arg \min \mathcal{X}_1^\mu(x) + \frac{1}{2\tau} \|x - \hat{x}\|^2$$



TOTAL VARIATION

$$\text{minimize} \quad \mu |\nabla x| + \frac{1}{2} \|x - f\|^2$$

Dualize the easy way...

$$|z| = \max_{\lambda \in [-1, 1]} z\lambda$$

$$\mu|z| = \max_{\lambda \in [-\mu, \mu]} z\lambda$$

$$\min \max_{\lambda \in [-\mu, \mu]} \langle \lambda, \nabla x \rangle + \frac{1}{2} \|x - f\|^2$$

why?

$$\max_{\lambda \in [-\mu, \mu]} \min_x \langle \nabla^T \lambda, x \rangle + \frac{1}{2} \|x - f\|^2$$

$$\nabla^T \lambda + x - f = 0$$

$$\langle \nabla^T \lambda, x \rangle$$

TOTAL VARIATION

$$\text{minimize} \quad \mu |\nabla x| + \frac{1}{2} \|x - f\|^2$$

$$\max_{\lambda \in [-\mu, \mu]} \min_x \langle \nabla^T \lambda, x \rangle + \frac{1}{2} \|x - f\|^2$$

optimality condition

$$\nabla^T \lambda + x - f = 0$$

$$x = f - \nabla^T \lambda$$

$$\max_{\lambda \in [-\mu, \mu]} \langle \nabla^T \lambda, f - \nabla^T \lambda \rangle + \frac{1}{2} \|\nabla^T \lambda\|^2$$

$$\max_{\lambda \in [-\mu, \mu]} \langle \nabla^T \lambda, f \rangle - \langle \nabla^T \lambda, \nabla^T \lambda \rangle + \frac{1}{2} \|\nabla^T \lambda\|^2$$

$$\max_{\lambda \in [-\mu, \mu]} \langle \nabla^T \lambda, f \rangle - \frac{1}{2} \|\nabla^T \lambda\|^2$$

$$\max_{\lambda \in [-\mu, \mu]} -\frac{1}{2} \|\nabla^T \lambda - f\|^2$$



FBS FOR TV

$$\text{minimize} \quad \mu |\nabla x| + \frac{1}{2} \|x - f\|^2$$

$$\max_{\lambda \in [-\mu, \mu]} -\frac{1}{2} \|\nabla^T \lambda - f\|^2 \quad \text{Dual} \quad x = f - \nabla^T \lambda$$

$$\mathcal{X}_\mu(z) = \begin{cases} 0, & z \in [-\mu, \mu] \\ \infty, & \text{otherwise} \end{cases}$$

$$\text{minimize} \quad \mathcal{X}_\mu(\lambda) + \frac{1}{2} \|\nabla^T \lambda - f\|^2$$

forward step

$$\hat{\lambda} = \lambda^k - \tau \nabla (\nabla^T \lambda^k - f)$$

backward step

$$\lambda^{k+1} = \arg \min \mathcal{X}_\mu(\lambda) + \frac{1}{2\tau} \|\lambda - \hat{\lambda}\|^2$$

$$= \min\{\max\{-\mu, \hat{\lambda}\}, \mu\}$$

SVM

$$\text{minimize} \quad \frac{1}{2} \|w\|^2 + Ch(YXw)$$

$$h(z) = \max\{1 - z, 0\}$$

use the trick

$$Ch(z) = \max_{\lambda \in [0, C]} \lambda(1 - z)$$

$$\min_w \max_{\lambda \in [0, C]} \frac{1}{2} \|w\|^2 + \langle \lambda, 1 - YXw \rangle$$

optimality $w - X^T Y \lambda = 0$

$$\text{maximize} \quad -\frac{1}{2} \|X^T Y \lambda\|^2 + \mathbf{1}^T \lambda$$

$$\text{subject to} \quad \lambda \in [0, C]$$

SVM

$$\begin{aligned} &\text{maximize} && -\frac{1}{2}\|X^T Y \lambda\|^2 + \mathbf{1}^T \lambda \\ &\text{subject to} && \lambda \in [0, C] \end{aligned}$$

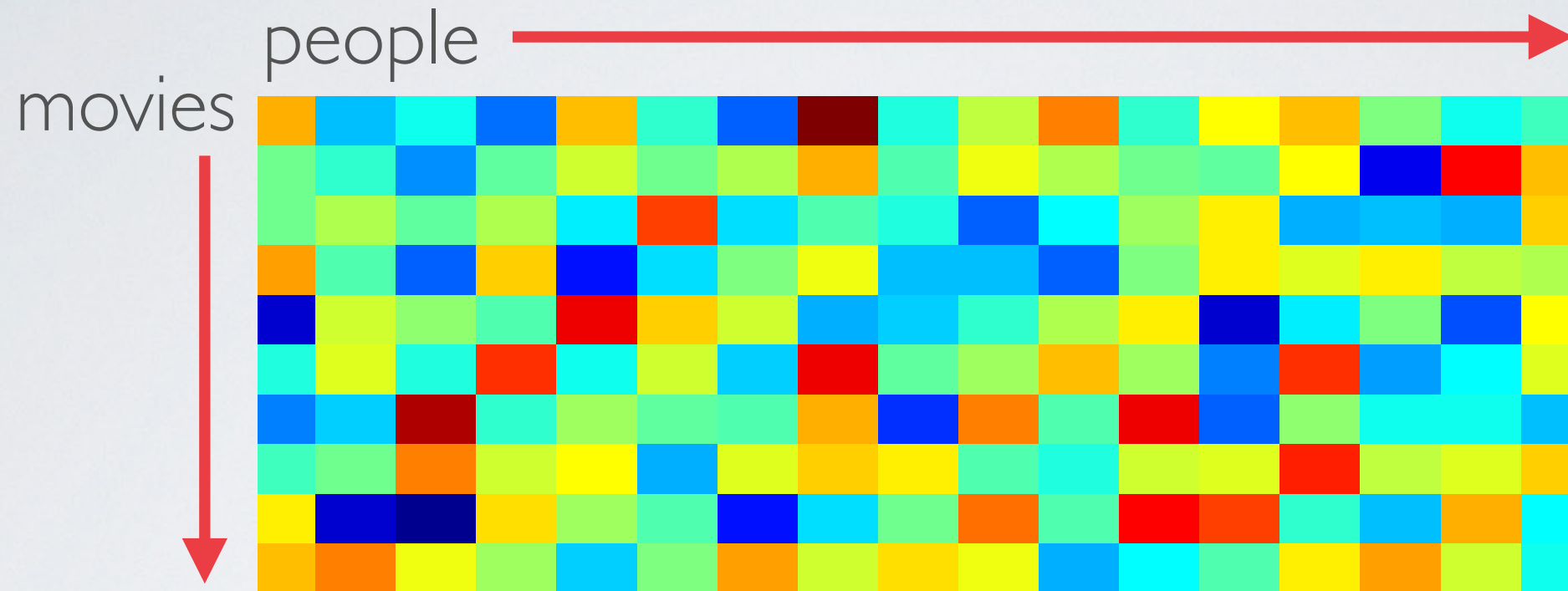
forward ASCENT step

$$\hat{\lambda} = \lambda^k - \tau(Y X X^T Y \lambda^k - \mathbf{1})$$

backward step

$$\lambda^{k+1} = \min\{\max\{0, \hat{\lambda}\}, C\}$$

NETFLIX PROBLEM

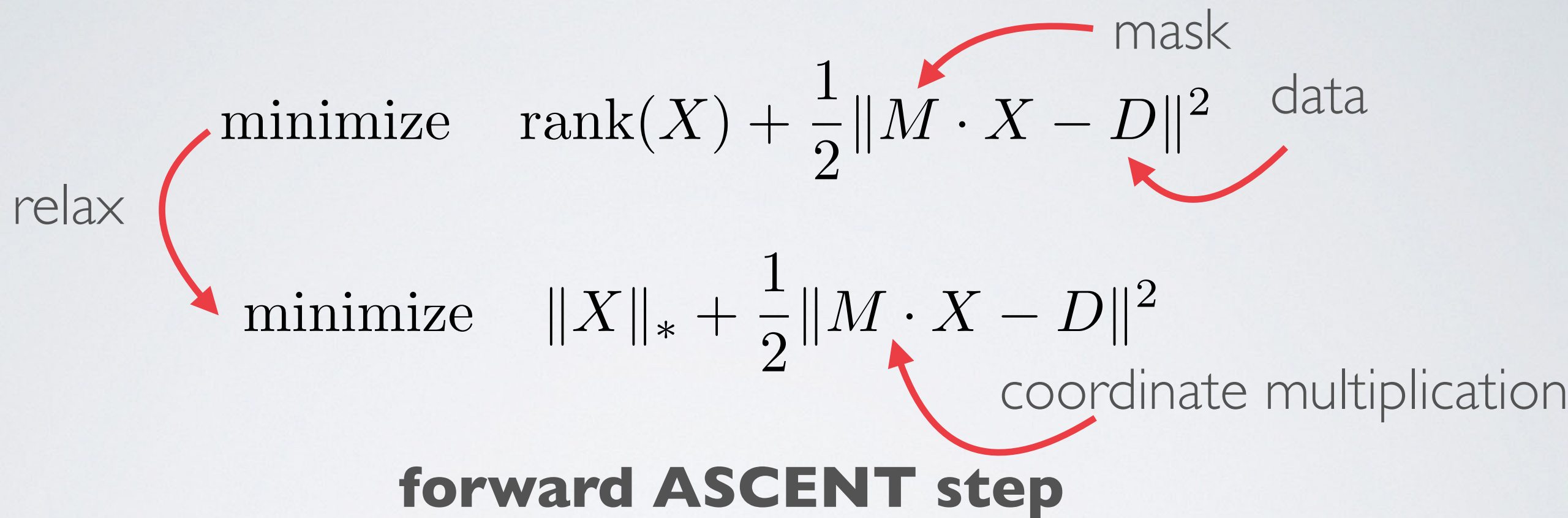


data “imputation”: fill in missing values

low rank assumption:
everyone described by affinity for horror, action, rom-com, etc...

“matrix completion”

NUCLEAR NORM FORM



$$\hat{X} = X^k - M \cdot (M \cdot X^k - D)$$

why not transpose?

backward step

$$X^{k+1} = \arg \min \|X\|_* + \frac{1}{2\tau} \|X - \hat{X}\|^2$$

what's this do?

MATRIX FACTORIZATION

$$\begin{aligned} & \underset{X, Y}{\text{minimize}} && \frac{1}{2} \|XY - D\|^2 \\ & \text{subject to} && X, Y \geq 0 \end{aligned}$$

forward step

$$\begin{aligned} \hat{X} &= X^k - \tau(X^k Y^k - D)(Y^k)^T \\ \hat{Y} &= Y^k - \tau(X^k)^T(X^k Y^k - D) \end{aligned}$$

backward step

$$\begin{aligned} X^{k+1} &= \max\{\hat{X}, 0\} \\ Y^{k+1} &= \max\{\hat{Y}, 0\} \end{aligned}$$

CONVERGENCE

by reduction to proximal-point method

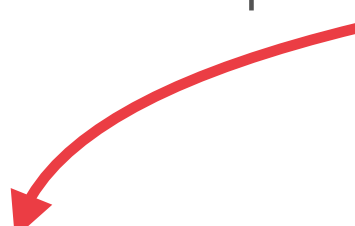
$$\text{minimize } h(x) = g(x) + f(x)$$

Theorem

Suppose the stepsize for FBS satisfies

$$\tau \leq \frac{2}{L}$$

does not
depend on g

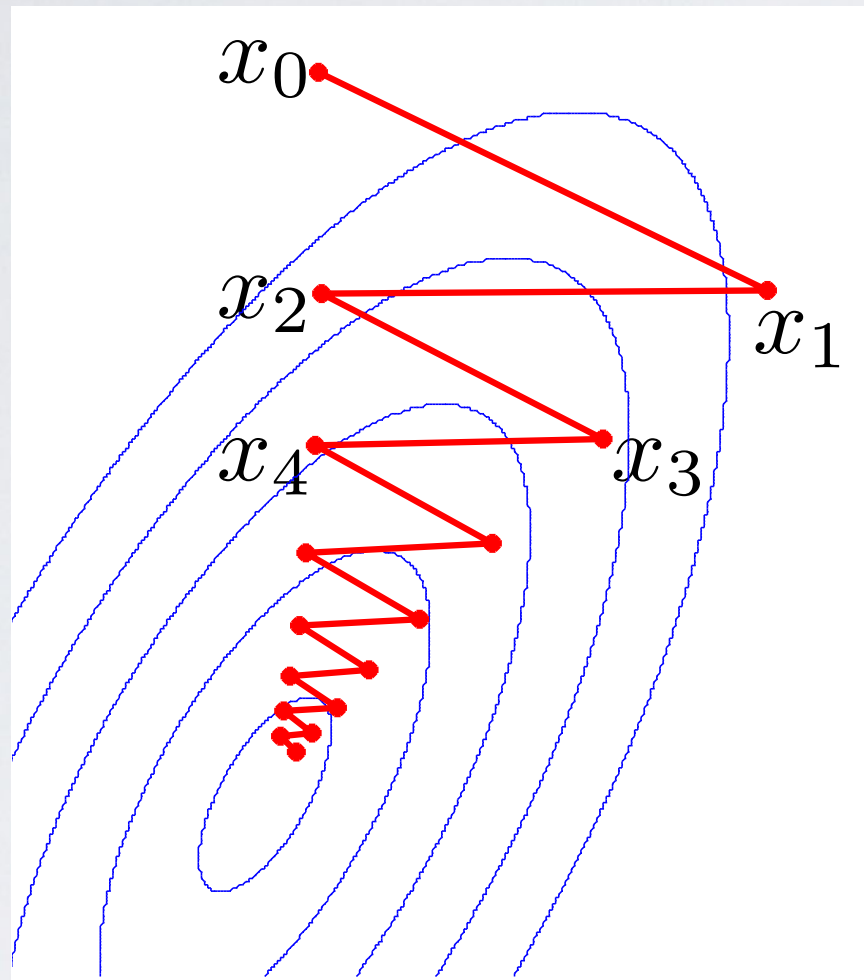


where L is a Lipschitz constant for ∇f . Then

$$h(x^k) - h(x^*) \leq O(1/k)$$

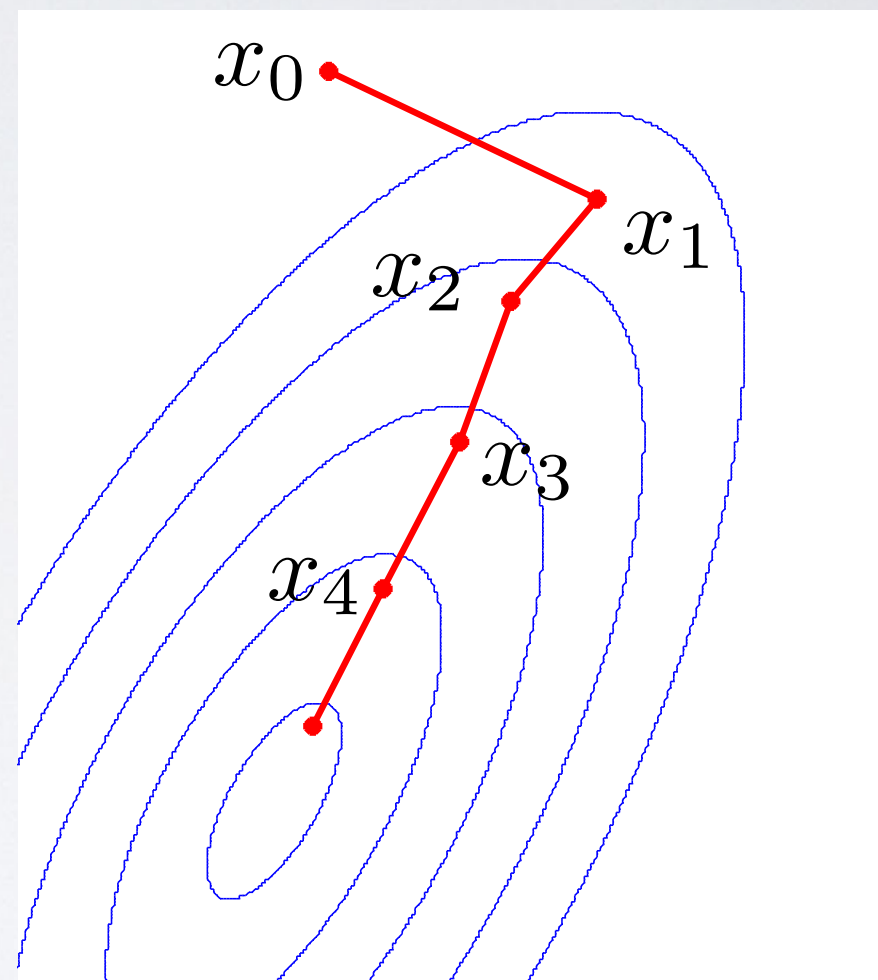
GRADIENT VS. NESTEROV

Gradient



$$O\left(\frac{1}{k}\right)$$

Nesterov



$$O\left(\frac{1}{k^2}\right) \leftarrow \text{Optimal}$$

FISTA

“Fast iterative shrinkage/thresholding” - Beck and Teboulle ‘09

FISTA

$$x^{k+1} = \text{prox}_g(y^k - \tau \nabla f(y^k), \tau)$$

$$\alpha^{k+1} = \frac{1}{2} \left(1 + \sqrt{4(\alpha^k)^2 + 1} \right)$$

$$y^{k+1} = x^{k+1} + \frac{\alpha^k - 1}{\alpha^{k+1}} \underbrace{(x^{k+1} - x^k)}_{\text{momentum}}$$

SPARSA / FASTA

idea: adaptive stepsize outperforms acceleration

$$f(x) \approx \frac{\alpha}{2} x^T x$$

secant equation

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

$$\Delta g = \alpha \Delta x$$

least-squares solution

$$\text{minimize} \quad \frac{1}{2} \|\alpha \Delta x - \Delta g\|^2$$

$$\alpha = \frac{\Delta x^T \Delta g}{\|\Delta x\|^2} \quad \longrightarrow \quad \tau = \frac{1}{\alpha} = \frac{\|\Delta x\|^2}{\Delta x^T \Delta g}$$

BACKTRACKING

based on MM interpretation

$$f(x^{k+1}) < f^k + \langle x^{k+1} - x^k, \nabla f(x^k) \rangle + \frac{1}{2\tau_k} \|x^{k+1} - x^k\|^2$$

behaves like Lipschitz constant



- Choose stepsize t_k FBS+Backtracking
- $x^{k+1} = \text{prox}_g(x^k - \tau_k \nabla f(x^k), \tau_k)$
- While $f(x^{k+1}) > f^k + \langle x^{k+1} - x^k, \nabla f(x^k) \rangle + \frac{1}{2\tau_k} \|x^{k+1} - x^k\|^2$

$$\tau_k \leftarrow \tau_k / 2$$

$$x^{k+1} = \text{prox}_g(x^k - \tau_k \nabla f(x^k), \tau_k)$$

STOPPING CONDITIONS

$$\text{minimize } h(x) = g(x) + f(x)$$

stop when “residual” is small....but what's the residual??

$$x^{k+1} = \arg \min g(x) + \frac{1}{2\tau} \|x - \hat{x}\|^2$$

$$0 \in \partial g(x^{k+1}) + \frac{1}{\tau} (x^{k+1} - \hat{x})$$

$$\frac{1}{\tau} (\hat{x} - x^{k+1}) \in \partial g(x^{k+1})$$

form the derivative

$$r^{k+1} = \frac{1}{\tau} \underbrace{(\hat{x} - x^{k+1})}_{\partial g(x^{k+1})} + \nabla f(x^{k+1}) \in \partial h(x^{k+1})$$

HOW SMALL IS SMALL?

RELATIVE RESIDUAL

$$\text{minimize } h(x) = g(x) + f(x)$$

residual

$$r^{k+1} = \frac{1}{\tau} \underbrace{(\hat{x} - x^{k+1})}_{\partial g(x^{k+1})} + \nabla f(x^{k+1}) \in \partial h(x^{k+1})$$

$$\text{stop when } r^{k+1} = \frac{1}{\tau}(\hat{x} - x^{k+1}) + \nabla f(x^{k+1}) \approx 0$$

$$\text{or equivalently } \frac{1}{\tau}(\hat{x} - x^{k+1}) \approx -\nabla f(x^{k+1})$$

relative residual

$$r_r^{k+1} = \frac{r^k}{\max\{\tau^{-1}(\hat{x} - x^{k+1}), \nabla f(x^{k+1})\}}$$