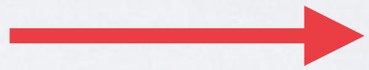
# PROXIMAL METHODS

# WHY PROXIMAL METHODS

Smooth functions

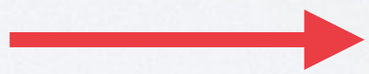
$$\text{minimize} \quad f(x)$$



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

Non-differentiable

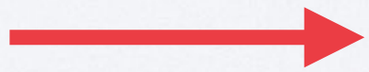
$$\text{minimize} \quad f(x)$$



**Proximal methods**

Constrained problems?

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g(x) \leq 0 \\ & h(x) = 0 \end{aligned}$$



Lagrangian methods

# PROXIMAL OPERATOR

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

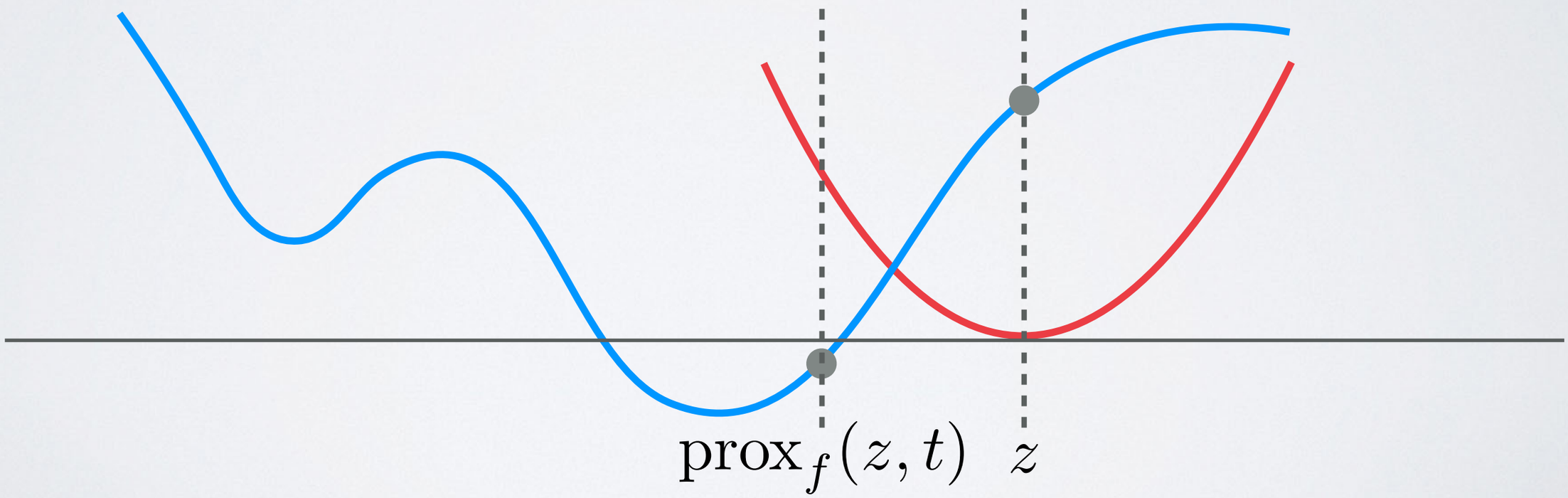
objective

proximal penalty
"Stay close to z"

# 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**

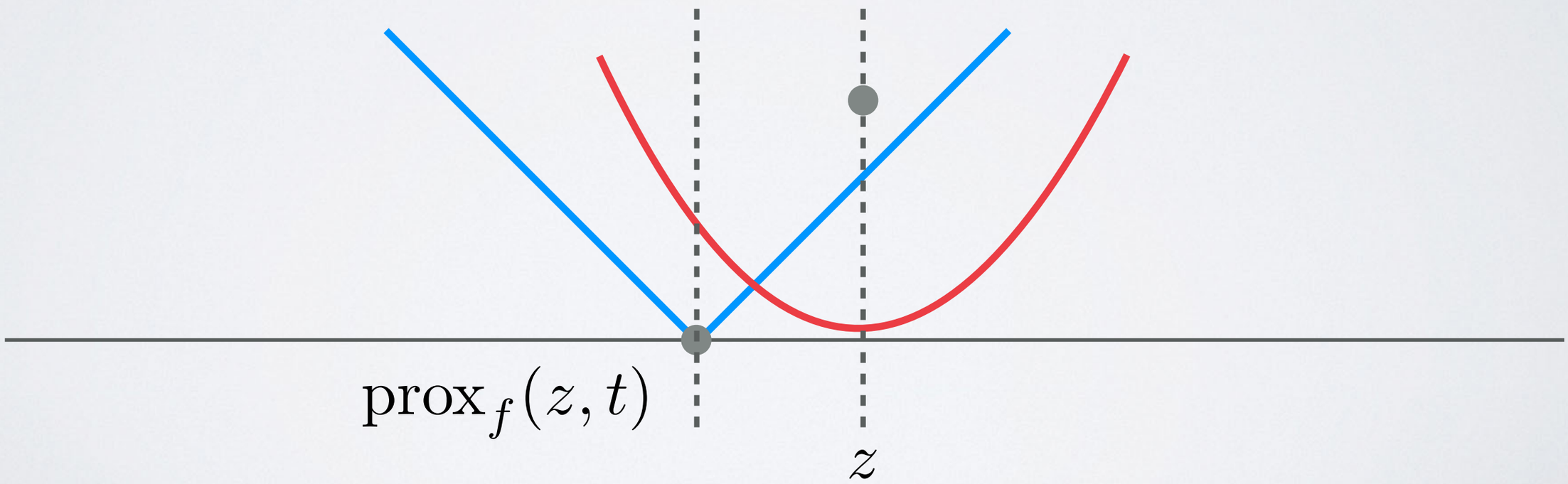
Forward gradient $\quad x = z - \tau \nabla f(z)$

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

# 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

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

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

forward gradient descent

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

**Existence**

Exists for any <u>proper</u> convex function (and some non-convex)

Exists only for smooth functions

**Uniqueness**

<u>Always</u> unique result

Sub-gradient unique if differentiable

# EXAMPLE: L1

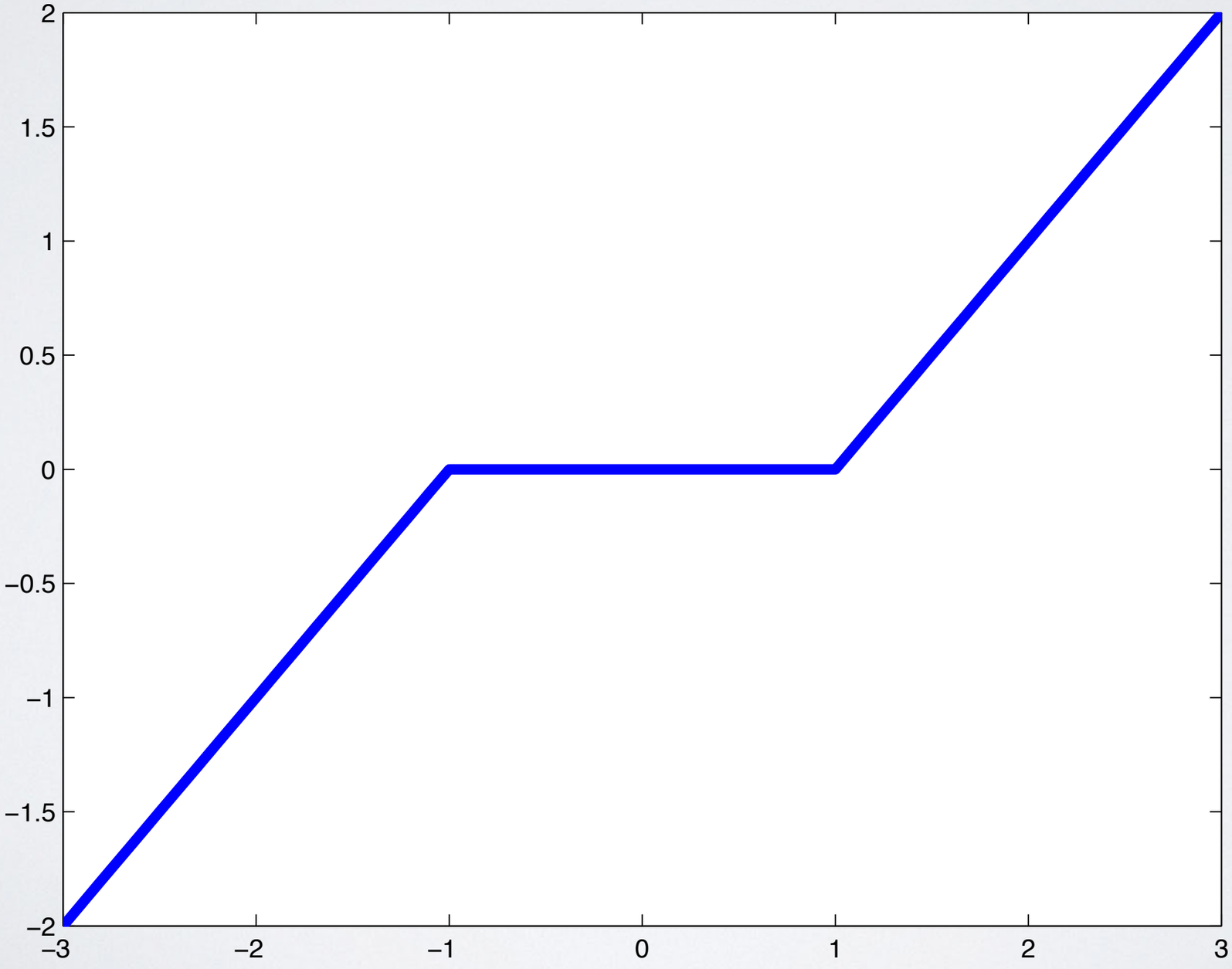
$$\text{prox}_f(z,\tau) = \arg\min_x |x| + \frac{1}{2\tau}\|x-z\|^2$$

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

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

# 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**"

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

# NUCLEAR NORM PROX

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

Same singular vectors

$$\text{minimize} \quad \|US_XV^T\|_* + \frac{1}{2\tau}\|US_XV^T - US_ZV^T\|^2$$

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

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

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

# CHARACTERISTIC FUNCTIONS

$$C = \text{some convex set}$$

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

proximal

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

What does this do?

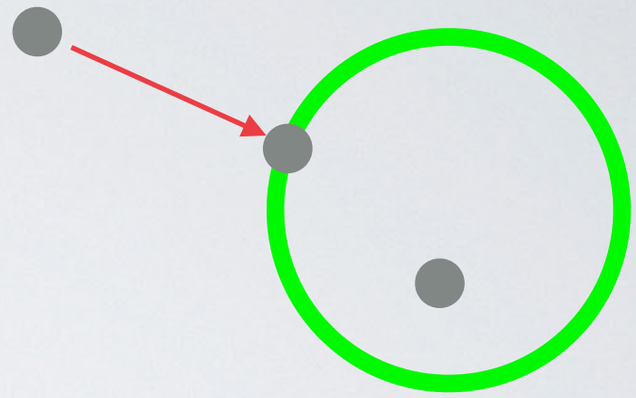
Does the stepsize matter?

# EXAMPLES

2-norm ball

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

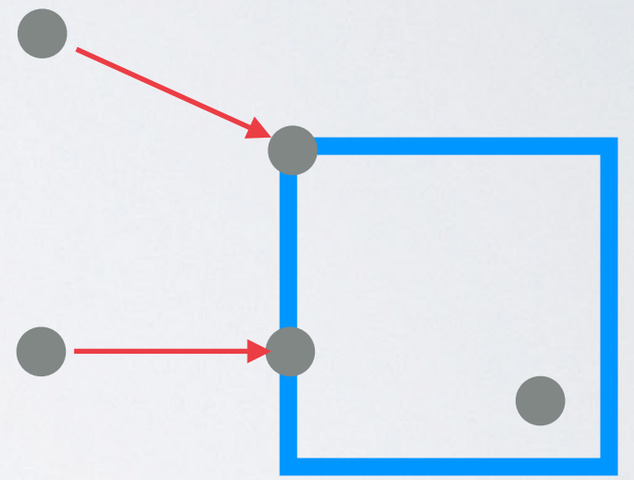
$$\text{prox}_2(z) = \frac{z}{\|z\|} \min\{\|z\|, 1\}$$



infinity ball

$$C = B_2 = \{x| \quad \|x\|_\infty \leq 1\}$$

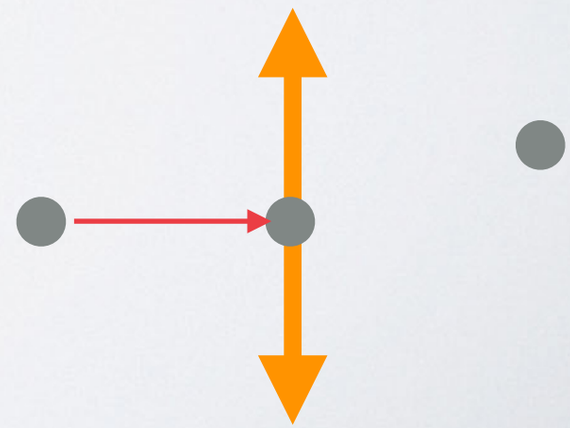
$$\text{prox}_\infty(z)_i = \min\{\max\{z_i, -1\}, 1\}$$



positive half space

$$C = \{x| \quad x \geq 0\}$$

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

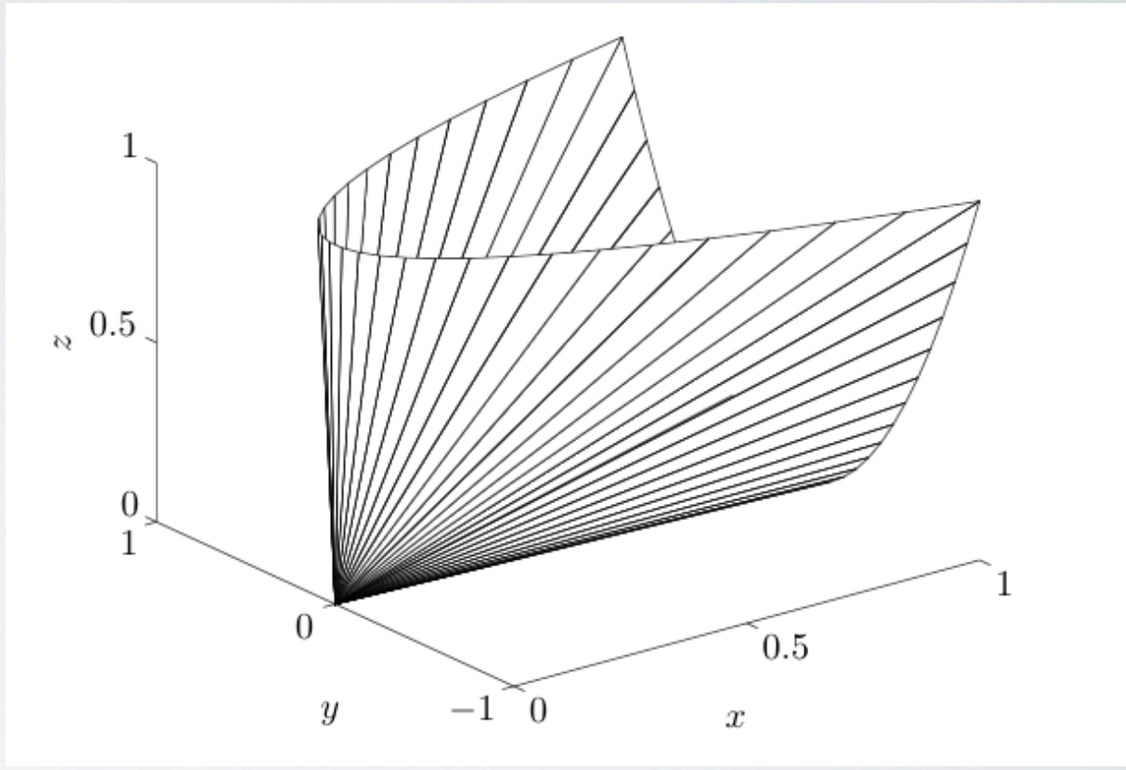


# HARDER EXAMPLES

semidefinite cone

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

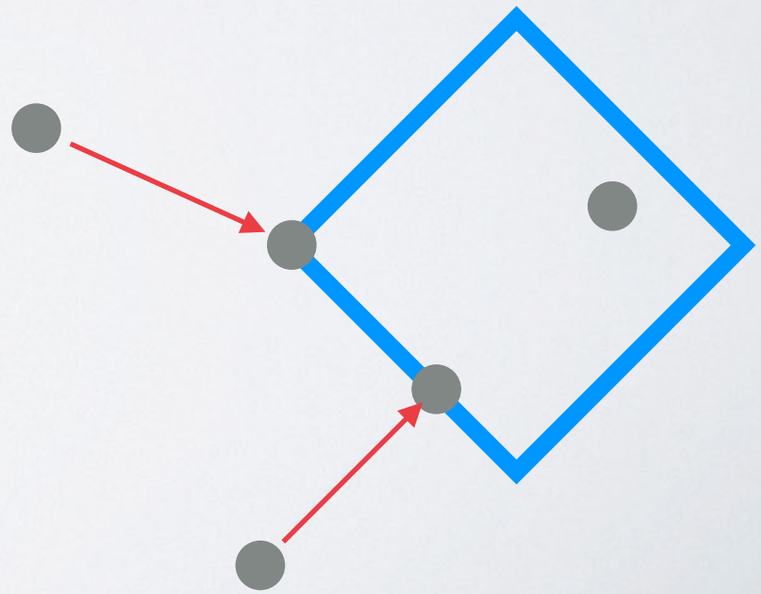
$$\text{prox}_{S_{++}}(Z) = U \max\{S, 0\} V^T$$




L1-ball

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

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



# PROXIMAL-POINT METHOD

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

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

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

# WHY FORWARD BACKWARD?

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

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

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

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

fixed-point property

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

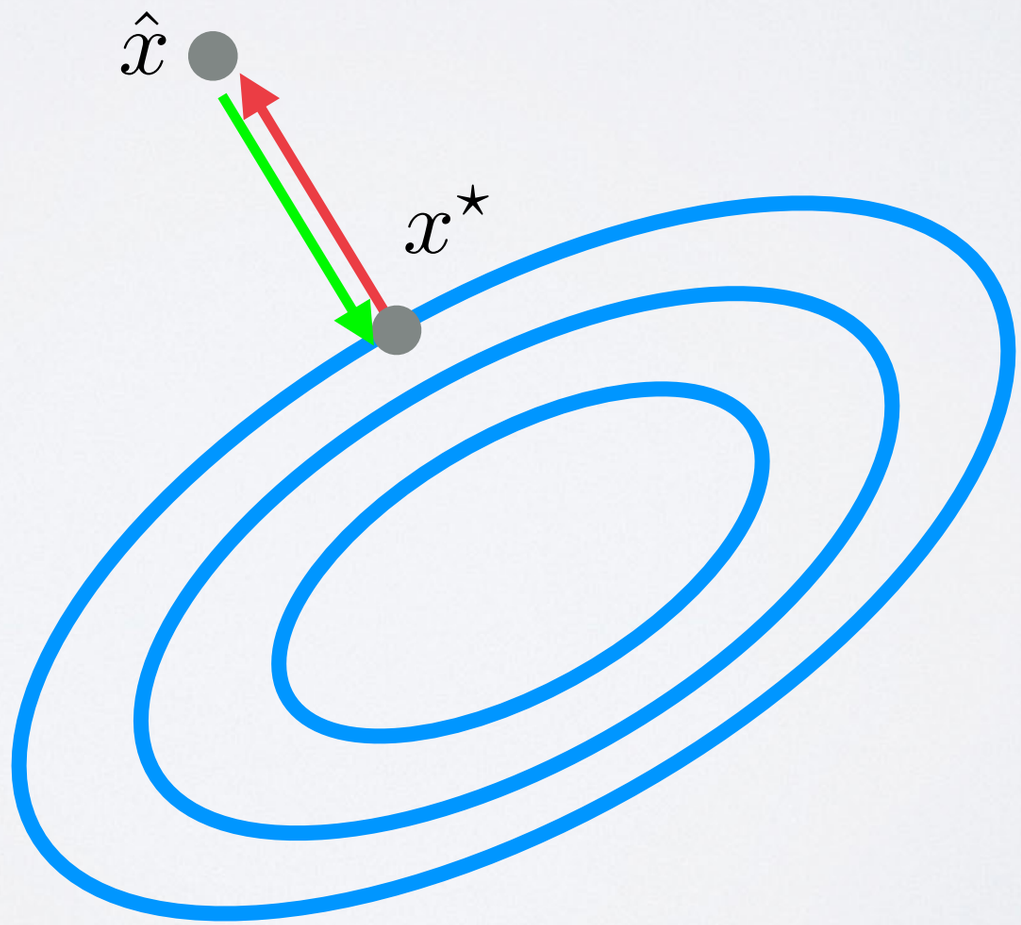
$$0 \in \nabla f(x^\star) + \partial g(x^\star)$$

# 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

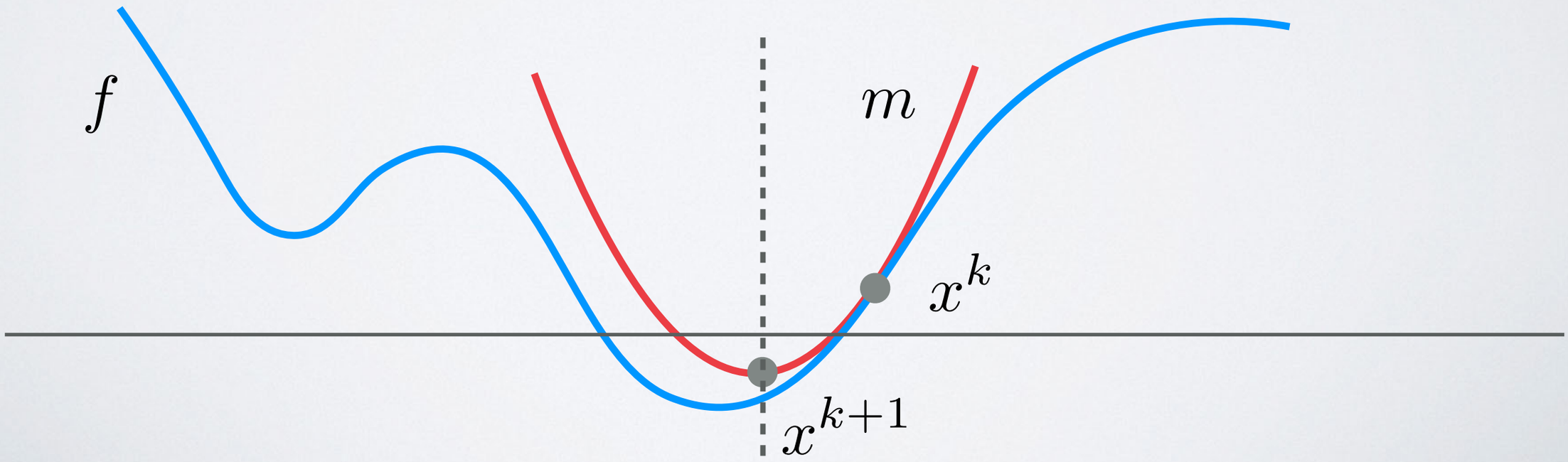
$$\text{minimize} \quad h(x)$$

**surrogate function**

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

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

surrogate "majorizes" f

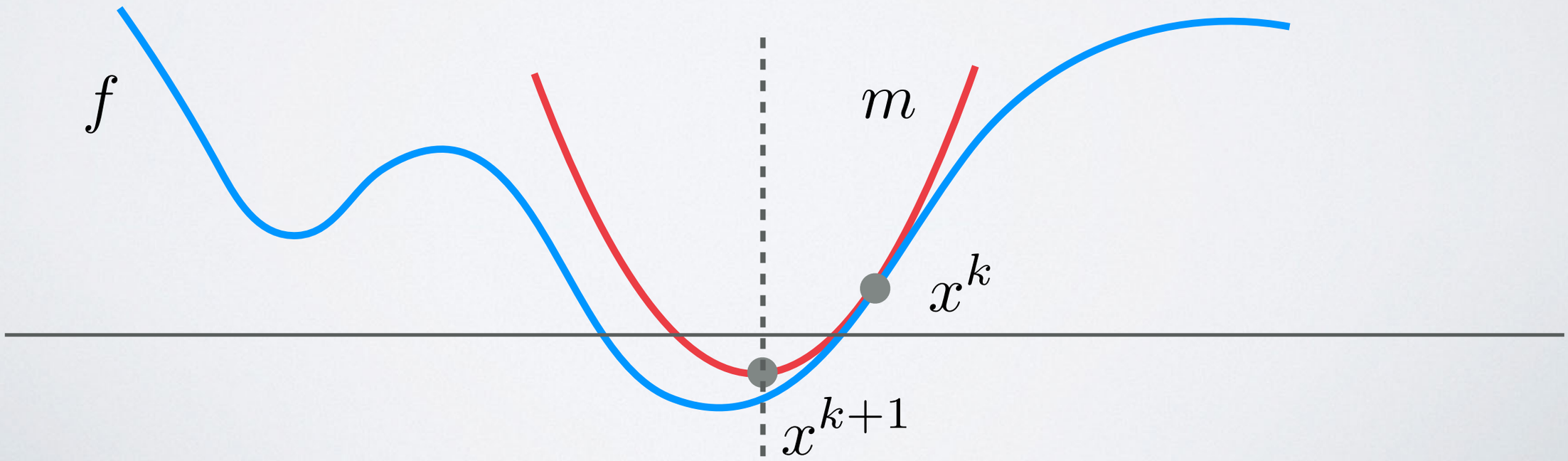


# MM ALGORITHM

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

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

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



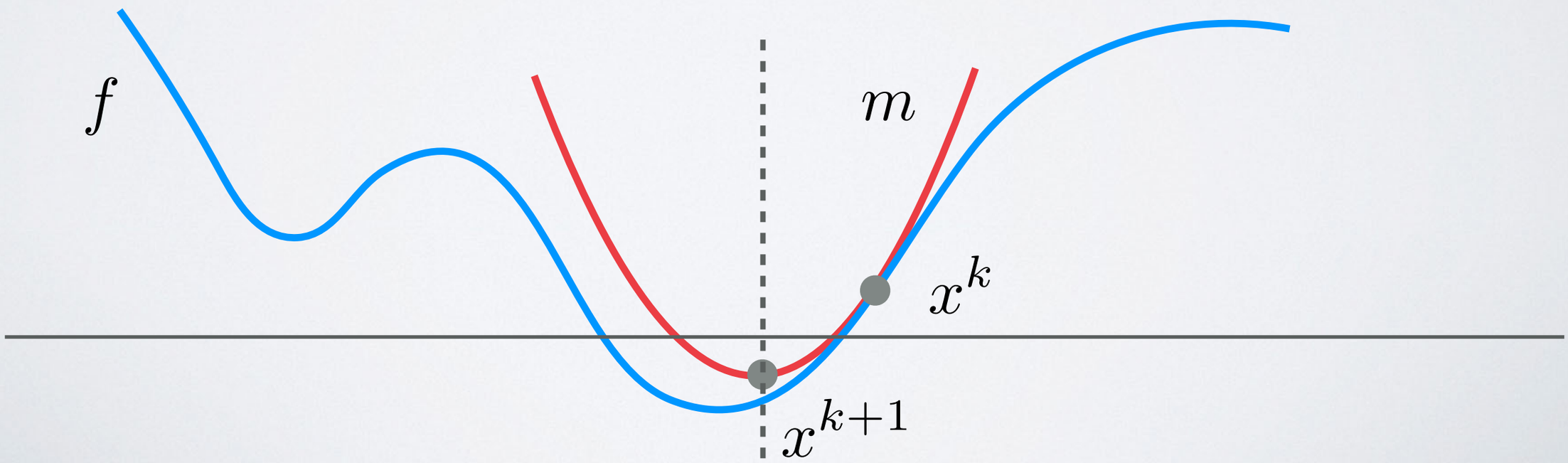
# MM ALGORITHM

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

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

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

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



# 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

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

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

non-smooth

smooth

## forward step

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

## backward step

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

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

iterative shrinkage / thresholding

# SPARSE LOGISTIC

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

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

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

## forward step

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

## backward step

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

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

# EXAMPLE: DEMOCRATIC REPRESENTATIONS

$$\begin{cases} \text{minimize} & \frac{1}{2}\|Ax-b\|^2 \\ \text{subject to} & \|x\|_\infty \le \mu \end{cases}$$

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

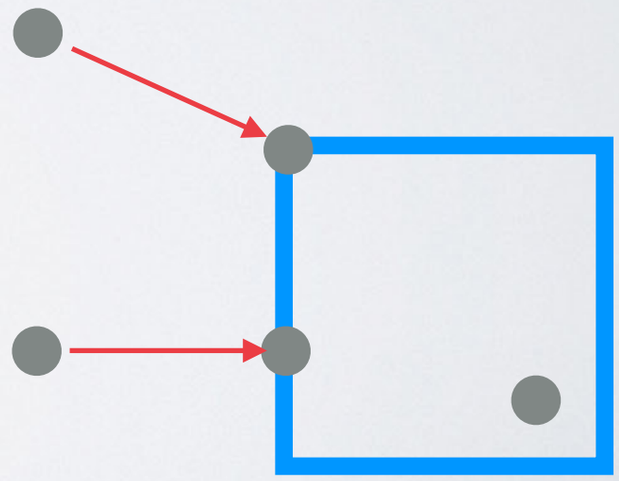
**forward step**

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

**backward step**

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

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



# EXAMPLE: LASSO

$$\begin{cases} \text{minimize} & \frac{1}{2}\|Ax-b\|^2 \\ \text{subject to} & |x| \le \mu \end{cases}$$

implicit constraints

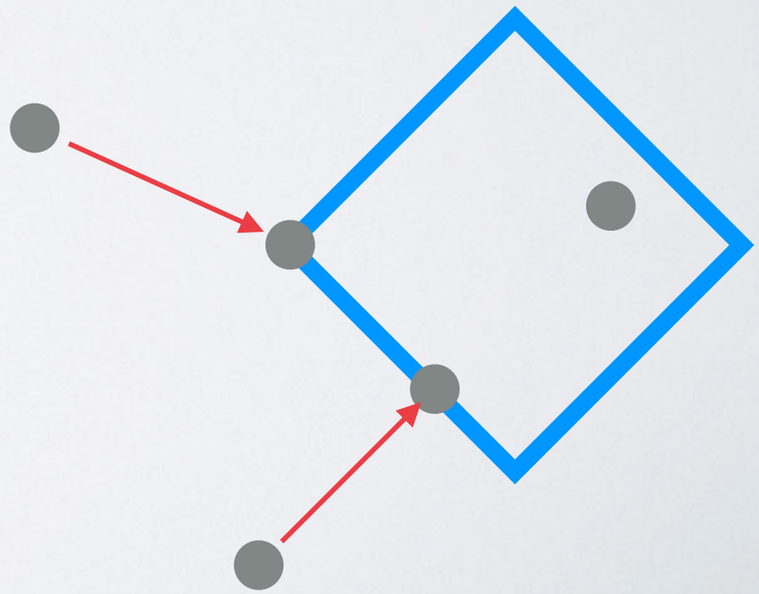
$$\text{minimize} \quad \mathcal{X}_1^\mu(x) + \frac{1}{2}\|Ax-b\|^2$$

**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…

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

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

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

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

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

# 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

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

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

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

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

**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 $\quad w - X^T Y \lambda = 0$

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

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

# SVM

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

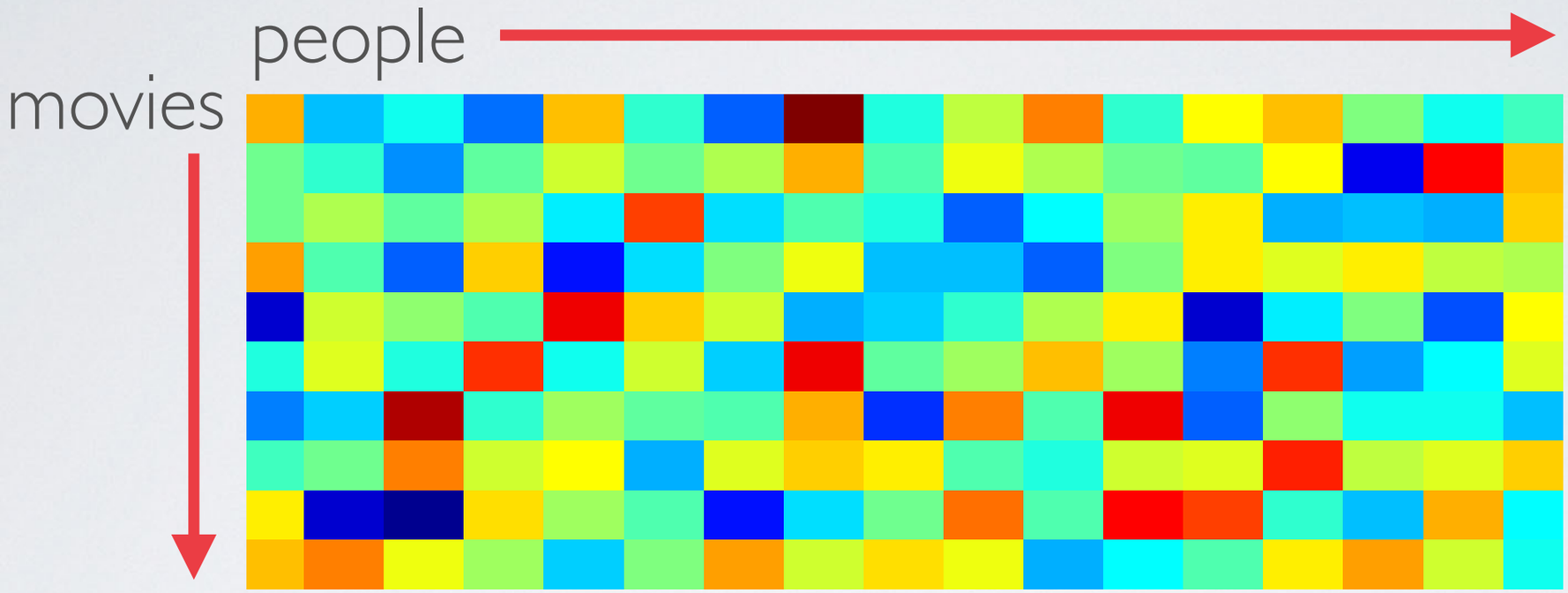
**forward ASCENT step**

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

**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

$$\text{minimize} \quad \text{rank}(X) + \frac{1}{2}\|M \cdot X - D\|^2$$

mask

data

relax

$$\text{minimize} \quad \|X\|_* + \frac{1}{2}\|M \cdot X - D\|^2$$

coordinate multiplication

**forward ASCENT step**

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

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

$$\underset{X,Y}{\text{minimize}} \quad \frac{1}{2}\|XY - D\|^2$$

$$\text{subject to} \quad X,\, Y \geq 0$$

**forward step**

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

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

**backward step**

$$X^{k+1} = \max\{\hat{X}, 0\}$$

$$Y^{k+1} = \max\{\hat{Y}, 0\}$$

# CONVERGENCE

by reduction to proximal-point method

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

**Theorem**

Suppose the stepsize for FBS satisfies does not depend on g

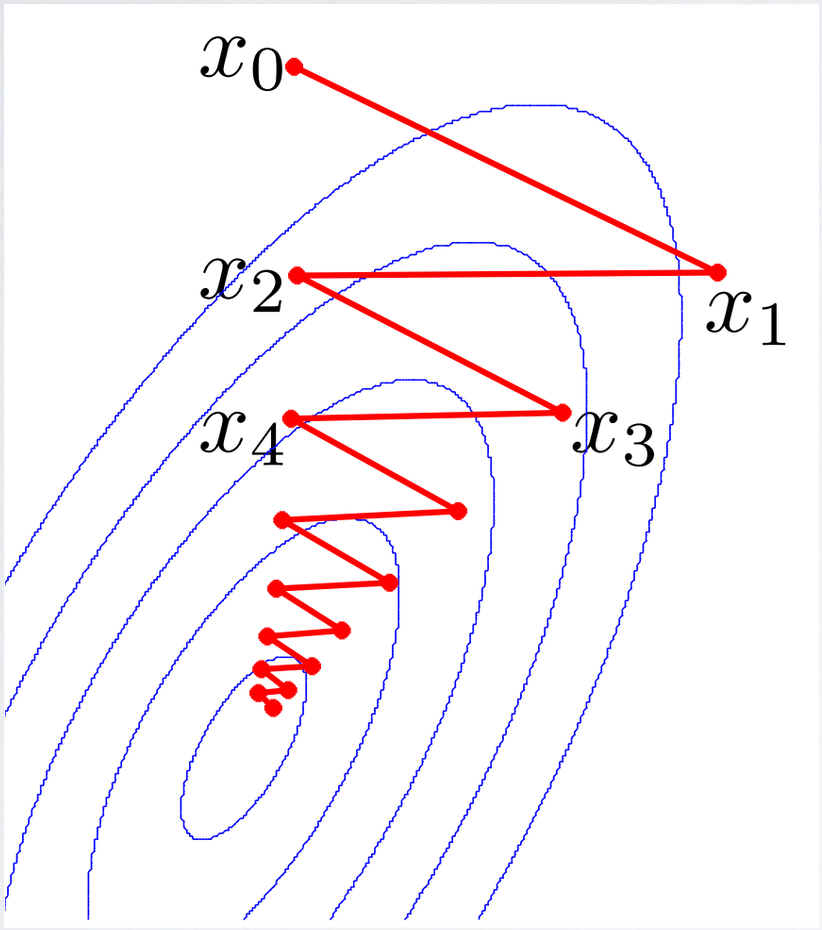
$$\tau \leq \frac{2}{L}$$

where L is a Lipschitz constant for $\nabla f$. Then

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

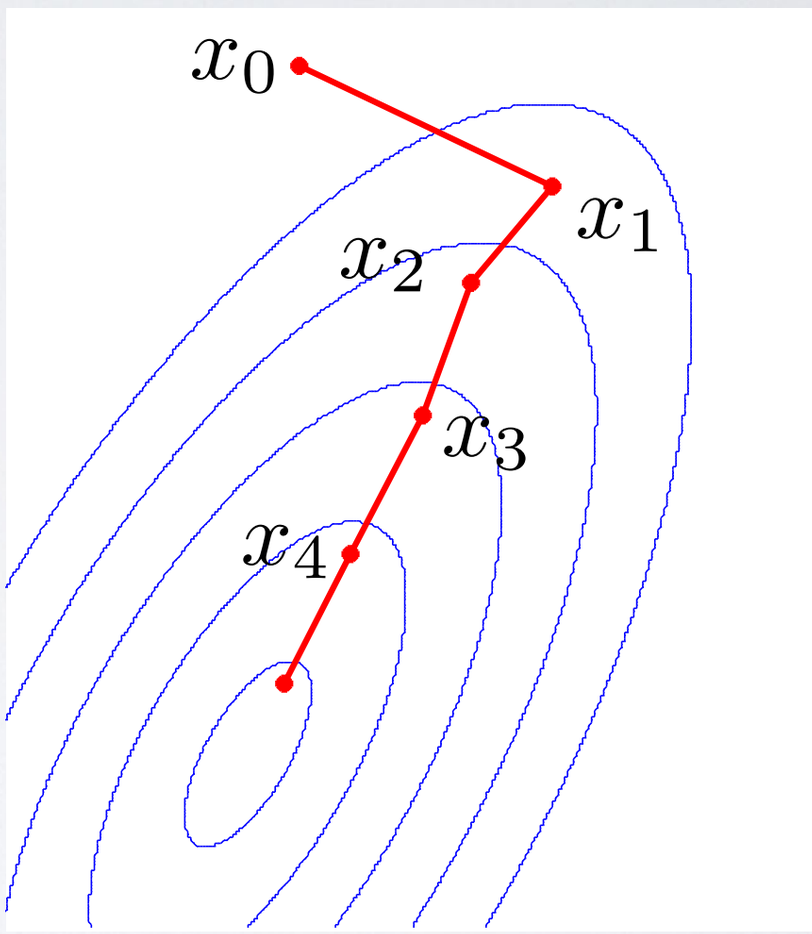
# GRADIENT VS. NESTEROV

## Gradient




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

## Nesterov




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

**Nemirovski and Yundin**

# FISTA

"Fast iterative shrinkage/thresholding" - Beck and Teboulle '09

**FISTA**

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

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

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

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

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

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

# 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

FBS+Backtracking

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

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

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

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

$$\partial g(x^{k+1})$$

# HOW SMALL IS SMALL?
# RELATIVE RESIDUAL

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

**residual**

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

$$\partial g(x^{k+1})$$

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

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

**relative residual**

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