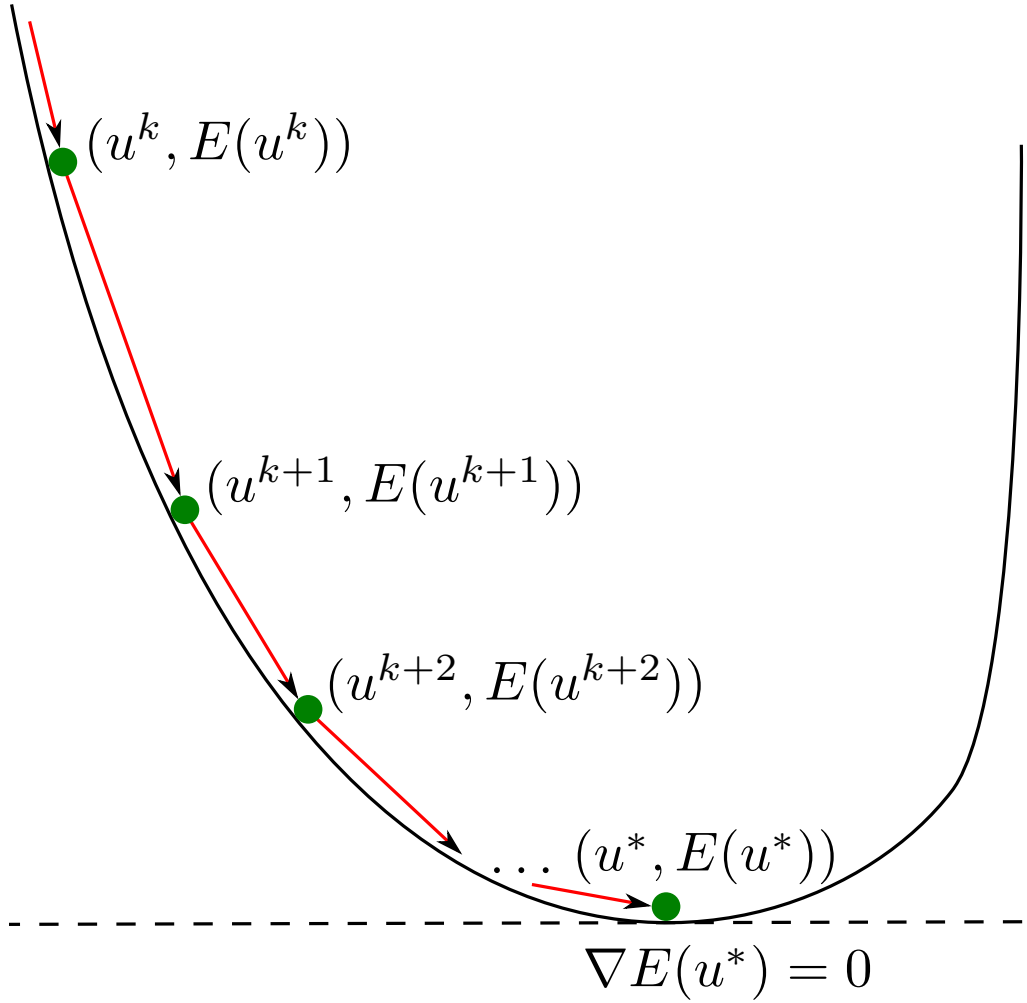
Gradient Methods

V. Estellers

WS 2017

Gradient Methods



Outline

Gradient Descent

Consider the unconstrained and smooth optimization problem

$$u^* \in \arg\min_{u \in \mathbb{R}^n} E(u),$$

for $E : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ proper, closed, and convex

Gradient descent is an optimization technique for the "simple" case

- dom $E = \mathbb{R}^n$
- $E \in \mathcal{C}^1(\mathbb{R}^n)$

Descent methods

Suppose we are at a point $u^k \in \mathbb{R}^n$ where $\nabla E(u^k) \neq 0$

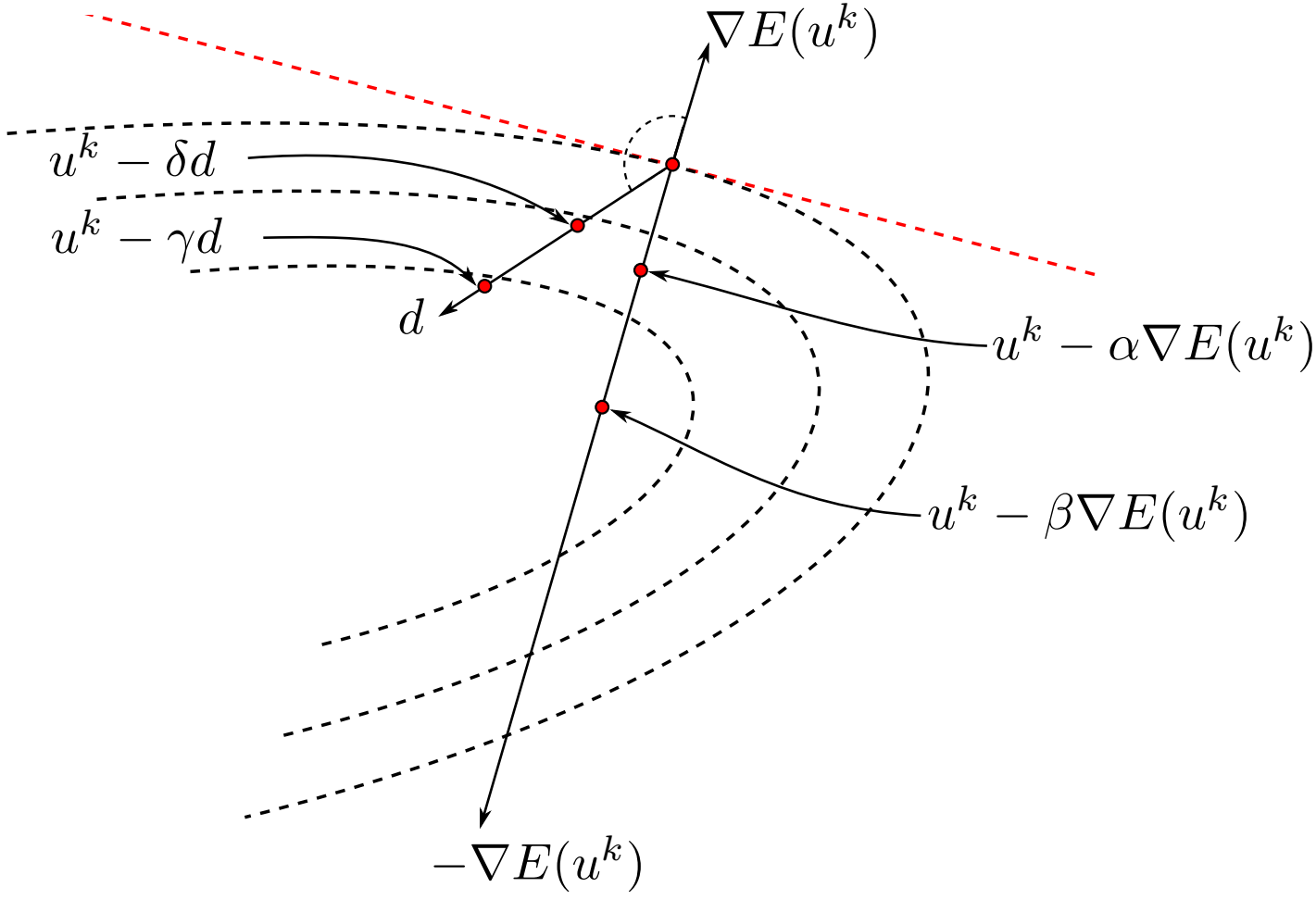
Consider the ray $u(\tau) = u^k + \tau d$ for some direction $d \in \mathbb{R}^n$

$$E(u(\tau)) = E(u^k + \tau d) = E(u^k) + \tau \langle \nabla E(u^k), d \rangle + o(\tau)$$

- $\tau \langle \nabla E(u^k), d \rangle$ dominates $o(\tau)$ for sufficiently small τ
- If $\langle \nabla E(u^k), d \rangle < 0$, d is a *descent direction* as, for suff. small τ,

$$E(u(\tau)) < E(u)$$

Descent methods



Descent methods

The negative gradient is the *steepest* descent direction

$$\underset{\|d\|=1}{\operatorname{argmin}} \ \left\{\langle d, \nabla E(u^k)\rangle\right\} = -\frac{\nabla E(u^k)}{\|\nabla E(u^k)\|}$$

The gradient is orthogonal to the iso-contours $\gamma : I \to \mathbb{R}^n$

$$\nabla E(\gamma(t)) \perp \dot{\gamma}(t), \qquad t \in I$$

Common choices of descent directions

- Scaled gradient: $d^k = -D^k \nabla E(u^k)$, $D^k \succeq 0$
- Newton: $D^k = [\nabla^2 E(u^k)]^{-1}$
- Quasi-Newton: $D^k \approx [\nabla^2 E(u^k)]^{-1}$
- Steepest descent: $D^k = I$

Gradient descent

Definition

Given a function $E \in \mathcal{C}^1(\mathbb{R}^n)$, an initial point $u^0 \in \mathbb{R}^n$ and a sequence $(\tau_k) \subset \mathbb{R}$ of step sizes, the iteration

$$u^{k+1} = u^k - \tau_k \nabla E(u^k), \qquad k = 0, 1, 2, \ldots,$$

is called *gradient descent*.

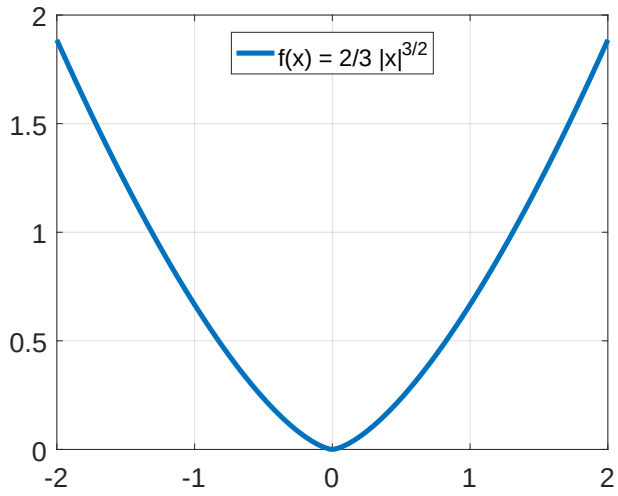
Philosophy:

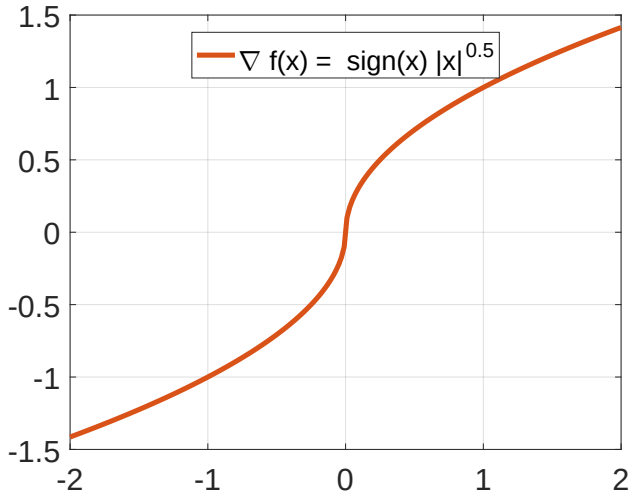
- Generate a decreasing sequence $\{E(u^k)\}_{k=0}^{\infty}$
- Each iteration is cheap, easy to code
- Choosing τ_k to guarantee convergence is not trivial

Constant step size

Consider a constant step size $\tau^k = \tau$

Will gradient descent work for any convex function?





For any constant time step $\tau > 0$, the starting point $u^0 = \left(\frac{\tau}{2}\right)^2$ results in a gradient descent sequence $u^0, -u^0, u^0, \ldots$

Intuition and requirements for constant step-size

Intuitively, an "infinitely quickly changing gradient" leads to "infinitely quickly changing" gradient descent updates

$$u^{k+1} = u^k - \tau_k \nabla E(u^k), \qquad k = 0, 1, 2, \ldots,$$

Need a stronger version of differentiability to prevent inf. quick changes

Definition: L-smooth function

If $E : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and its first derivative is Liptschitz continuous, i.e. there exists an $L \geq 0$ such that

$$\|\nabla E(u) - \nabla E(v)\| \leq L \|u - v\|, \forall u, v \in \mathbb{R}^n,$$

then E is called L-smooth

Lipschitz continuity

Reminder

$f : \mathbb{R}^n \to \mathbb{R}^m$ is called Lipschitz continuous if for some $L \geq 0$

$$\|f(x) - f(y)\| \leq L \|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

If the function is differentiable, we can characterize Lipschitz continuous functions by the size of its gradient.

Theorem: Lipschitz continuity for differentiable functions

A differentiable function $E : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz with parameter L if and only if $\|\nabla E(x)\|_{S^\infty} \leq L$ for all $x \in \mathbb{R}^n$.

Convergence Analysis

Conjecture

For any L-smooth proper convex function E (with a minimizer) there exists a step size τ such that the gradient descent algorithm converges

To prove this conjecture, we will use a general **fixed-point Iteration** for algorithms of the form

$$u^{k+1} = G(u^k)$$

Example:

$$G(u) = u - \tau \nabla E(u).$$

If the iteration converges to $\hat{u}$ and ∇E is continuous, then $\nabla E(\hat{u}) = 0$.

Outline

Convergence of Fixed-Point Iterations

References:

Ryu and Boyd, *Primer on Monotone Operator Methods*, 2016.

Burger, Sawatzky, and Steidl, *First Order Algorithms in Variational Image Processing*, 2017.

Bauschke, and Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2011.

Fixed-point iterations with contractions

When does the fixed-point iteration

$$u^{k+1} = G(u^k) \tag{1}$$

converge?

Banach fixed-point theorem

If the update rule $G : \mathbb{R}^n \to \mathbb{R}^n$ is a **contraction**, i.e. if there exists a $L < 1$ such that

$$\|G(u) - G(v)\|_2 \leq L\|u - v\|_2$$

holds for all $u,\ v \in \mathbb{R}^n$, then the iteration (1) converges to the unique fixed-point $\hat{u}$ of G. More precisely,

$$\|u^k - \hat{u}\|_2 \leq L^k \|u^0 - \hat{u}\|_2.$$

Fixed-point iterations with averaged operators

G being a **contraction** is **too restrictive** in many cases

G being **non-expansive**, i.e. Lipschitz continuous with constant $L = 1$, is commonly true.

- any rotation G is non-expansive and has a fixed point (0)
- the iteration $u^{k+1} = G(u^k)$ does not converge

Averaged operator

An operator $G : \mathbb{R}^n \to \mathbb{R}^n$ is called **averaged** if there exists a non-expansive mapping $H : \mathbb{R}^n \to \mathbb{R}^n$ and a constant $\alpha \in (0, 1)$ such that

$$G = \alpha I + (1 - \alpha)H.$$

Criteria for being averaged

Lemma about nonexpansive operators

Convex combinations as well as compositions of nonexpansive operators are nonexpansive.

Being averaged for smaller α

If a function $G : \mathbb{R}^n \to \mathbb{R}^n$ is averaged with respect to $\alpha \in]0, 1[$, then it is also averaged with respect to any other parameter $\tilde{\alpha} \in]0, \alpha[$.

Composition of averaged operators

If $G_1 : \mathbb{R}^n \to \mathbb{R}^n$ and $G_2 : \mathbb{R}^n \to \mathbb{R}^n$ are averaged, then $G_2 \circ G_1$ is also averaged.

Proofs: Notes

Criteria for being averaged

Firmly non-expansive

A function $G : \mathbb{R}^n \to \mathbb{R}^n$ is called **firmly nonexpansive**, if for all $u, v \in \mathbb{R}^n$ it holds that

$$\|G(u) - G(v)\|_2^2 \leq \langle G(u) - G(v), u - v \rangle.$$

Firmly nonexpansive operators are averaged

A function $G : \mathbb{R}^n \to \mathbb{R}^n$ is firmly nonexpansive if and only if G is averaged with $\alpha = \frac{1}{2}$.

Proof: Notes

Convergence for averaged operators

Krasnosel'skii-Mann Theorem

If the operator $G : \mathbb{R}^n \to \mathbb{R}^n$ is averaged and has a fixed-point, then the iteration

$$u^{k+1} = G(u^k)$$

converges to a fixed point of G for any starting point $u^0 \in \mathbb{R}^n$.

Proof: Notes

Short summary

We have seen:

An operator G is called a **contraction** if it is Lipschitz continuous with $L < 1$.

Contractions have a unique fixed-point and their fixed-point iteration converges with $\mathscr{O}(L^k)$.

An operator R is called a **nonexpansive** if it is Lipschitz continuous with $L = 1$.

An operator G is called a **averaged** if $G = \alpha I + (1 - \alpha)R$ for some nonexpansive operator R and $\alpha \in (0, 1)$.

If an averaged operator has a fixed-point, then the fixed-point iteration converges. The convergence rate states that $\sum_{k=1}^{n} \|G(u^k) - u^k\|_2 \leq C$ for some constant C.

Firmly nonexpansive operators are the same as averaged operators with $\alpha = \frac{1}{2}$.

Relation to gradient descent

We now have two loose ends:

- a conjecture about the convergence of the gradient descent iteration
- theorem that states the convergence of a fixed-point iteration for averaged operators.

we need to write gradient descent as an averaged operator

Baillon-Haddad theorem

A continuously differentiable convex function $E : \mathbb{R}^n \to \mathbb{R}$ is L-smooth if and only if $\frac{1}{L}\nabla E$ is firmly nonexpansive, i.e.

$$\langle \nabla E(u) - \nabla E(v), u - v \rangle \geq \frac{1}{L} \|\nabla E(u) - \nabla E(v)\|_2^2$$

for all $u, v \in \mathbb{R}^n$.

Proof: See Nesterov, *Introductory Lectures on Convex Optimization*, Theorem 2.1.5.

Outline

Convergence of gradient descent

Gradient descent as an averaged operator

If $E : \mathbb{R}^n \to \mathbb{R}$ has a minimizer, is convex and L-smooth, and $\tau \in]0, \frac{2}{L}[$, then the gradient descent iteration converges to a minimizer.

Sufficient: $G(u) = u - \tau \nabla E(u)$ is averaged.

We know $\frac{1}{L}\nabla E$ is averaged with $\alpha = 1/2$, i.e., $\frac{1}{L}\nabla E = \frac{1}{2}(I + T)$ for a non-expansive T.

It hold that

$$G(u) = u - \tau L \frac{1}{L}\nabla E(u) = \left(1 - \frac{L\tau}{2}\right) I + \frac{L\tau}{2}(-T)$$

If T is non-expansive, $(-T)$ is non-expansive, too.
$\Rightarrow$ For $\tau \in]0, \frac{2}{L}[$, G is averaged.

Convergence rate

How fast does gradient descent converge?

Theory of averaged operators shows $\sum_k \|\nabla E(u^k)\|_2^2$ is bounded.

Careful analysis shows that for L-smooth functions with $\tau \in (0, \frac{2}{L})$:

$$E(u^{k+1}) \leq E(u^k) \qquad E(u^k) - E(u^*) \in \mathcal{O}(1/k)$$

.

It is not possible to get a contraction to speed up convergence because a contraction would imply the existence of a unique fixed-point.

Reminder

$\mathcal{O}(g) = \{f \mid \exists C \geq 0, \exists n_0 \in \mathbb{N}_0, \forall n \geq n_0 : |f(n)| \leq C|g(n)|\}$

Strongly-convex + L-smooth

Gradient descent as an averaged operator

If $E : \mathbb{R}^n \to \mathbb{R}$ is m-strongly convex and L-smooth, and $\tau \in]0, \frac{2}{m+L}[$, then the gradient descent iteration converges to the unique minimizer u^* of E with $\|u^k - u^*\| \leq c^k \|u^0 - u^*\|$.

Proof on the Notes.

Strong convexity

Definition: strong convexity

A function $E : \mathbb{R}^n \to \overline{\mathbb{R}}$ is called *strongly convex* with constant m or m-strongly convex if $E(u) - \frac{m}{2}\|u\|_2^2$ is still convex.

Theorem: characterization of m-strongly convex functions [1]

For $E \in \mathcal{C}^1(\mathbb{R}^n)$ the following are equivalent:

1. $E(u) - \frac{m}{2}\|u\|^2$ is convex
2. $E(v) \geq E(u) + \langle \nabla E(u), v - u \rangle + \frac{m}{2}\|v - u\|^2$
3. $\langle \nabla E(u) - \nabla E(v), u - v \rangle \geq m\|u - v\|^2$
4. $\nabla^2 E(u) \succeq m \cdot I$, if $E \in \mathcal{C}^2(\mathbb{R}^n)$

[1] Ryu, Boyd, A Primer on Monotone Operator Methods, Appendix A

Back to GD

Optimal convergence rates

In computer vision, m-strongly convex L-smooth energies are very rare! Can one do better than the $\mathcal{O}(1/k)$ in the L-smooth case?

Famous analysis by Nesterov, (Th 2.1.7 and Th2.1.13) for first order methods of the form:

$$u^{k+1} \in u^0 + \text{span}\{\nabla E(u^0), \ldots, \nabla E(u^k)\}$$

If E can be any convex L-smooth function

then no first order method can have a worst-case complexity less than $\mathcal{O}(1/k^2)$.

and E is m-strongly convex, then no first order method can have a worst-case complexity less than $\mathcal{O}((\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^{2k})$ for $\kappa = L/m$.

Obtaining optimal convergence rates

Nesterov's Accelerated Gradient Descent

Pick some starting point $v^0 = u^0$, and iterate

1. Compute
$$u^{k+1} = v^k - \frac{1}{L}\nabla E(v^k)$$
2. Find the next $\alpha \in]0, 1[$ by solving
$$\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 + \frac{m}{L}\alpha_{k+1}$$
3. Compute the extrapolation of u^{k+1} via
$$\beta_k = \frac{\alpha_k(1 - \alpha_k)}{\alpha_k^2 + \alpha_{k+1}}$$
$$v^{k+1} = u^{k+1} + \beta_k(u^{k+1} - u^k)$$

Backtracking line search

Sometimes Lipschitz constant L not known

The convergence analysis shows that one really only needs

$$E(u^{k+1}) \leq E(u^k) - \beta_k \|\nabla E(u^k)\|^2$$

for some $\beta_k \geq \beta > 0$.

Idea: Pick $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$

Then determine τ_k each iteration by:

$\tau_k \leftarrow 1$

while $E\left(u^k - \tau_k \nabla E(u^k)\right) > E(u^k) - \alpha\tau_k \left\|\nabla E(u^k)\right\|^2$

$\tau_k \leftarrow \beta\tau_k$

end

Backtracking line search

Line search...

- ... often leads to improved convergence in practice
- ... has a (slight) overhead each iteration
- ... has the same convergence rate as with constant steps

For a backtracking line search scheme for Nesterov's accelerated gradient method please see *Introductory Lectures on Convex Optimization*, page 76, scheme (2.2.6).

Remark: Other strategies for linear search exists, e.g.

$$\tau_k = \arg\min_{\tau} E(u^k - \tau \nabla E(u^k))$$

Application: TV image denoising

Lets consider the applications of image denoising:

Application: TV image denoising

Lets consider the applications of image denoising:

Via energy minimization: Let D_1 and D_2 be finite difference operators for the partial derivatives. Determine

$$\hat{u} \in \arg\min_u \underbrace{\frac{\lambda}{2}\|u-f\|_2^2}_{\text{How close to input}} + \underbrace{\sum_{x\in\Omega}\sqrt{(D_1u(x))^2+(D_2u(x))^2}}_{\text{Total variation}}$$

Application: TV image denoising

Problem: The so called *total variation regularization*

$$TV(u) = \sum_{x \in \Omega} \sqrt{(D_1 u(x))^2 + (D_2 u(x))^2}$$

is not differentiable!

Application: TV image denoising

Problem: The so called *total variation regularization*

$$TV(u) = \sum_{x \in \Omega} \sqrt{(D_1 u(x))^2 + (D_2 u(x))^2}$$

is not differentiable!

Idea: Approximate it with a differentiable function

$$TV_\epsilon(u) = \sum_{x \in \Omega} \phi \sqrt{(D_1 u(x))^2 + (D_2 u(x))^2 + \epsilon^2}$$

Application: TV image denoising

Problem: The so called *total variation regularization*

$$TV(u) = \sum_{x \in \Omega} \sqrt{(D_1 u(x))^2 + (D_2 u(x))^2}$$

is not differentiable!

Idea: Approximate it with a differentiable function

$$TV_\epsilon(u) = \sum_{x \in \Omega} \phi \sqrt{(D_1 u(x))^2 + (D_2 u(x))^2 + \epsilon^2}$$

Exercises: Our denoising model is L-smooth for

$$L = \lambda + \frac{\|D\|_{S^\infty}}{\epsilon}$$

Application: TV image denoising

Problem: The so called *total variation regularization*

$$TV(u) = \sum_{x \in \Omega} \sqrt{(D_1 u(x))^2 + (D_2 u(x))^2}$$

is not differentiable!

Idea: Approximate it with a differentiable function

$$TV_\epsilon(u) = \sum_{x \in \Omega} \phi \sqrt{(D_1 u(x))^2 + (D_2 u(x))^2 + \epsilon^2}$$

Exercises: Our denoising model is L-smooth for

$$L = \lambda + \frac{\|D\|_{S^\infty}}{\epsilon}$$

We expect the convergence to be better for large ϵ, but we expect $TV(u) \approx TV_\epsilon(u)$ only for small ϵ...

Image denoising

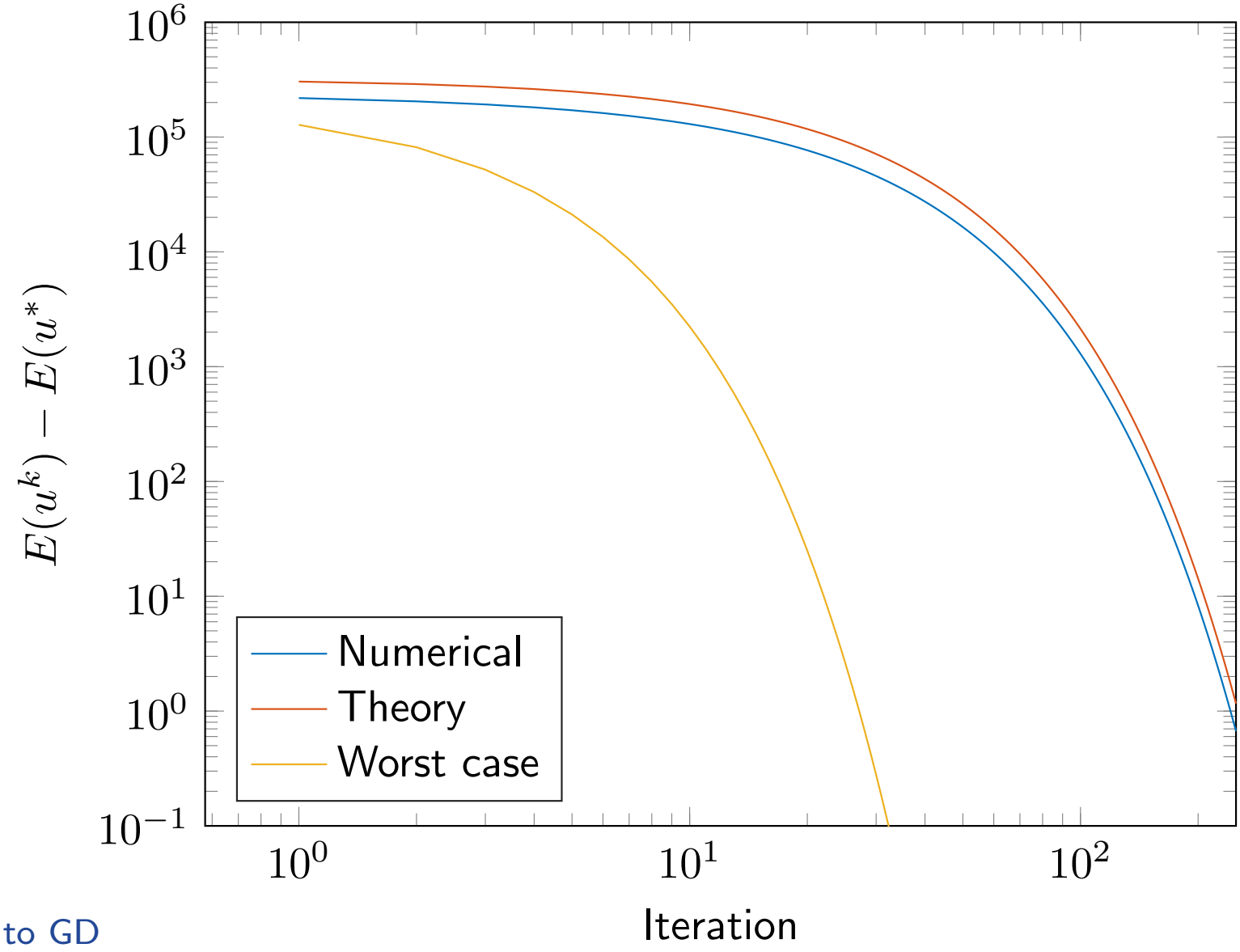
$\varepsilon = 0.1$

$\varepsilon = 0.01$

Back to GD

$\rightarrow$ *Motivation for non-smooth optimization!*

Convergence, $\tau = 2/(m + L)$



Convergence, backtracking line search

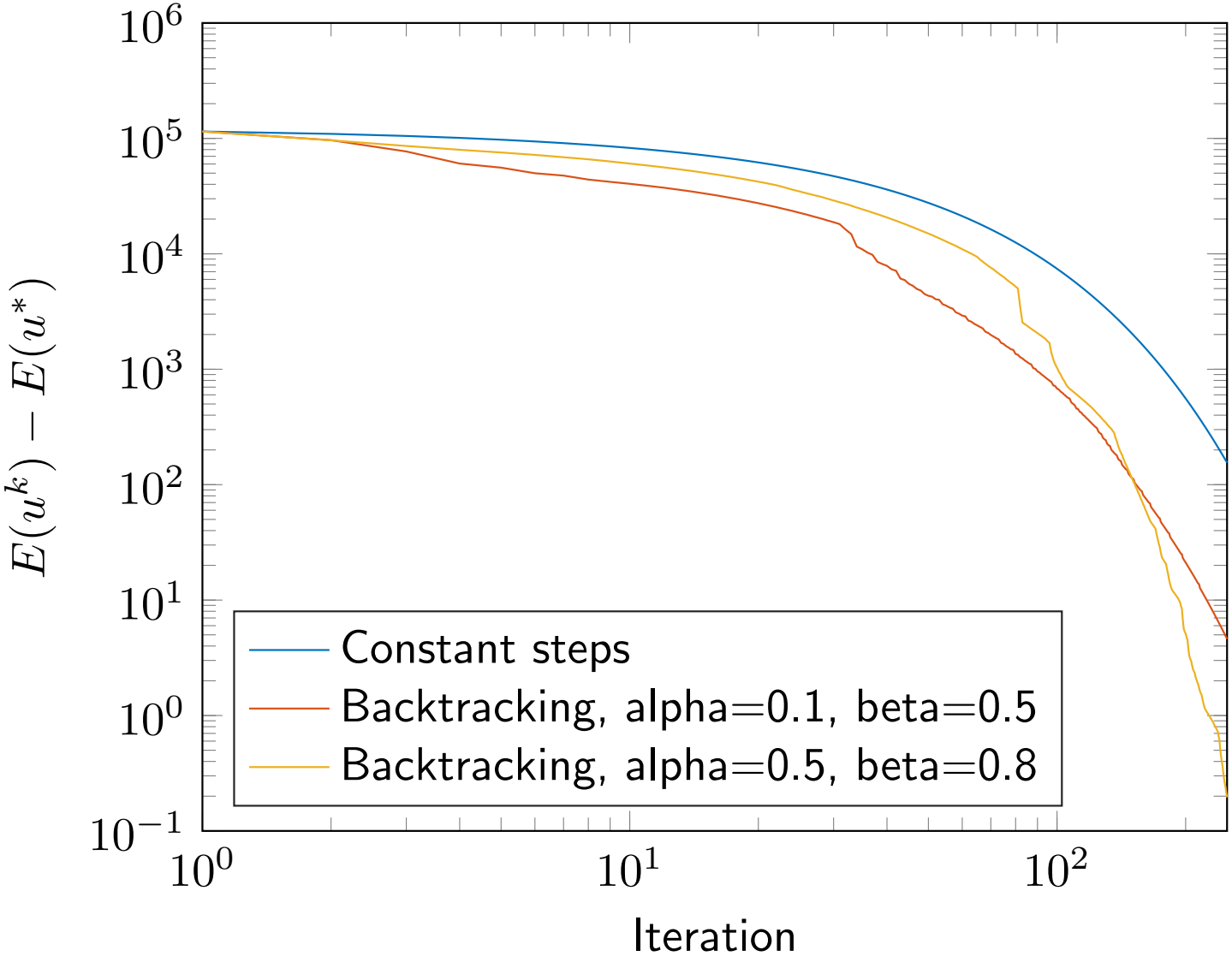
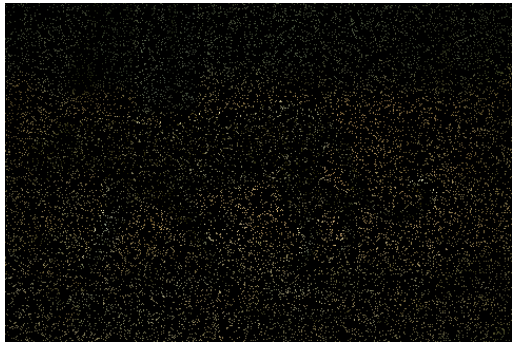
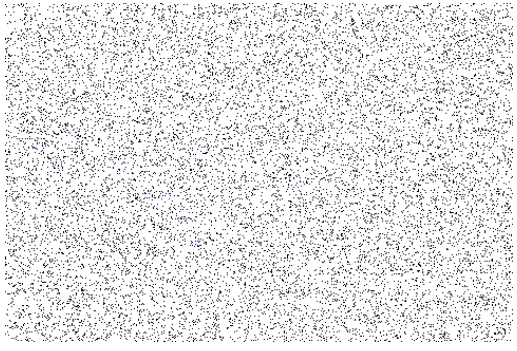


Image inpainting



$f \in \mathbb{R}^N$



$1 - m \in \mathbb{R}^N$

$u^* \in \mathbb{R}^N$

$$u^* \in \operatorname*{argmin}_u \frac{\lambda}{2} \|m \cdot (u - f)\|^2 + TV_\epsilon(u)$$

Energy is not strongly convex, but L-smooth

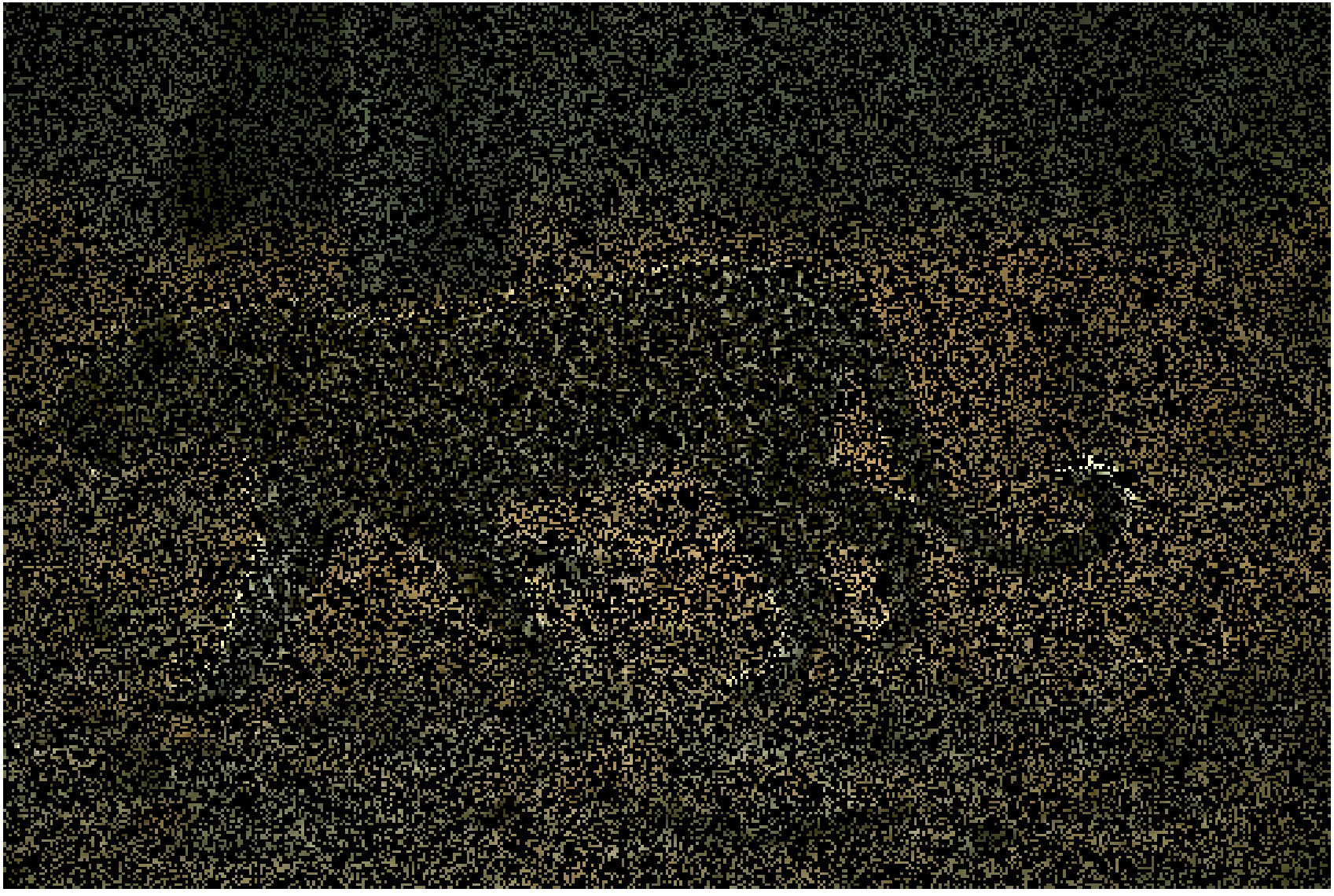
Sublinear upper bound on convergence speed

Image Inpainting

50% missing pixels

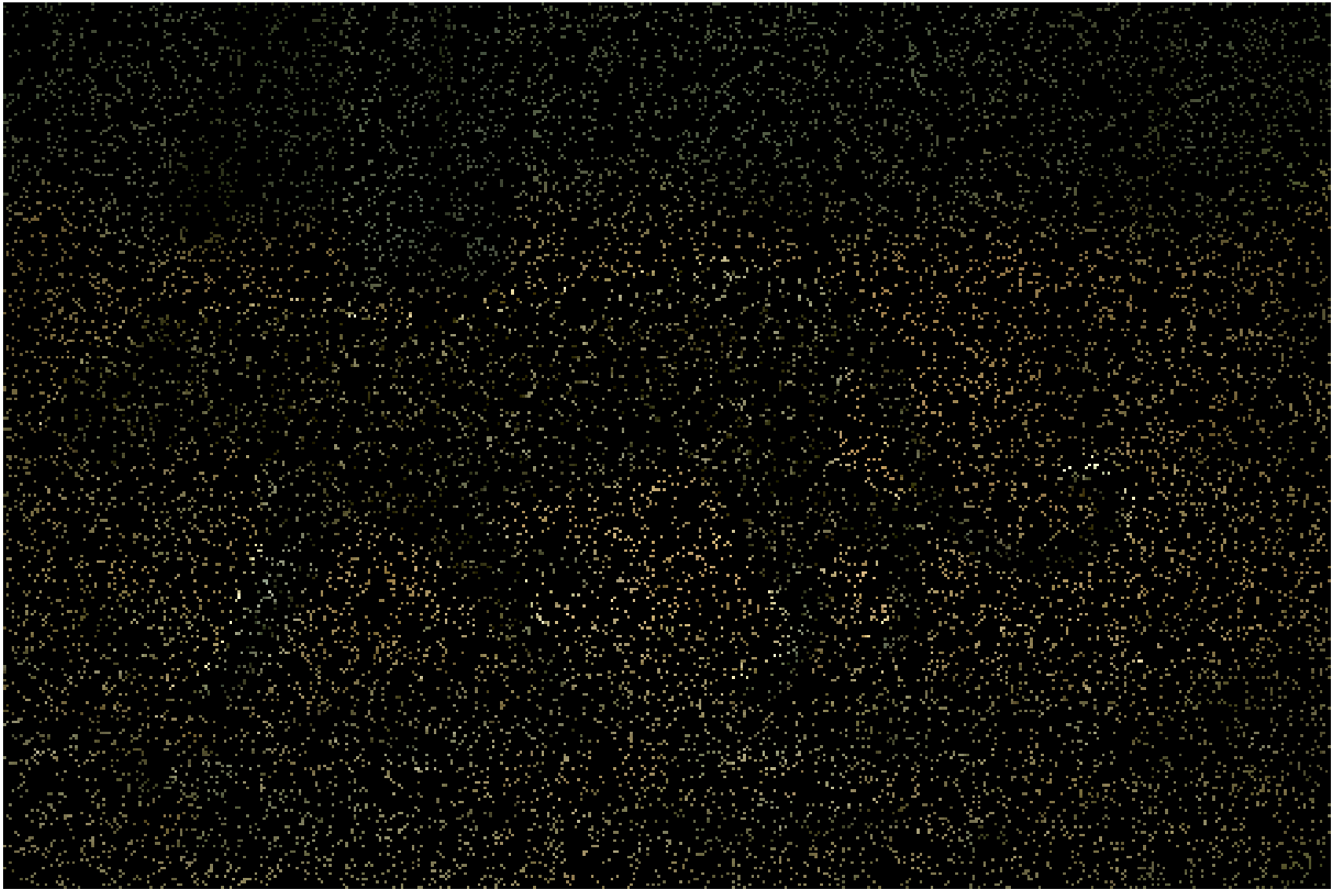
50% missing pixels

70% missing pixels



70% missing pixels

90% missing pixels



90% missing pixels

Concluding remarks and outlook

GD is still popular to date due to its simplicity and flexibility

Various theoretically optimal extensions (Heavy-ball acceleration, Nesterov momentum) exist

Envelope approach: many advanced algorithms for non-smooth optimization are just gradient descent on a particular (albeit complicated) energy

Endless of variants and modifications of descent methods

conjugate, accelerated, preconditioned, projected, conditional, mirrored, stochastic, coordinate, continuous, online, variable metric, subgradient, proximal, ...

Subgradient descent in one slide

We have seen in the exercises, that even for functions that are not L-smooth, gradient descent with a small step size reduces the energy up to some point where it starts oscillating.
Possible convergent variant: **Subgradient descent**

$$u^{k+1} = u^k - \tau_k p^k, \qquad \text{for any } p^k \in \partial E(u^k).$$

If it holds that

- E has a minimizer
- E is Lipschitz continuous
- $\tau_k \to 0$, but $\sum_{k=1}^{n} \tau_k \to \infty$, e.g. $\tau_k = 1/k$

then the subgradient descent iteration converges with

$$E(u^k) - E(u^*) \in \mathcal{O}(1/\sqrt{k})$$

Summary

This lecture is about

$$u^* \in \arg\min_{u \in \mathbb{R}^n} E(u),$$

for $E : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ proper, closed, convex.

Gradient descent:

dom $E = \mathbb{R}^n$

For L-smooth E (that has a minimizer)

- energy convergence in $\mathcal{O}(1/k)$ for constant step sizes
- energy convergence in $\mathcal{O}(1/k^2)$ for Nesterov's method.

For L-smooth m-strongly convex E: energy and iterate convergence in $\mathcal{O}(c^k)$

Summary

This lecture is about

$$u^* \in \arg\min_{u \in \mathbb{R}^n} E(u),$$

for $E : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ proper, closed, convex.

Gradient descent:

dom $E = \mathbb{R}^n$

For L-smooth E (that has a minimizer)

- energy convergence in $\mathcal{O}(1/k)$ for constant step sizes
- energy convergence in $\mathcal{O}(1/k^2)$ for Nesterov's method.

For L-smooth m-strongly convex E: energy and iterate convergence in $\mathcal{O}(c^k)$

Line search strategies for unknown Lipschitz constant L.