CPSC 661: Sampling Algorithms in ML

Andre Wibisono

March 15, 2021

Yale University

Last time

- Optimization
- Gradient flow in continuous time
- Strong convexity $\Rightarrow$ Gradient dominated $\Rightarrow$ Sufficient growth
- Exponential convergence to minimizer

Today: Algorithms for optimization

Optimization

$$\min_{x \in \mathcal{X}} f(x)$$

Gradient flow:

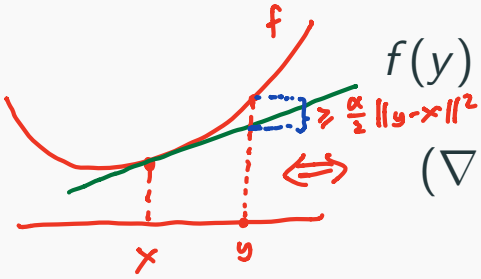
$$\dot{X}_t = -\nabla f(X_t)$$

Today assume $\mathcal{X} = \mathbb{R}^n$, but also for manifold

Strong convexity

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and $\min f = \min_{x \in \mathbb{R}^n} f(x)$

1. f is α-strongly convex if



$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|x - y\|^2$$

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \alpha \|x - y\|^2$$

$$\Leftrightarrow \quad \nabla^2 f(x) \succeq \alpha I$$

$$\Downarrow$$

2. f is α-gradient dominated if

$$\|\nabla f(x)\|^2 \geq 2\alpha \left(f(x) - \min f\right)$$

(also known as Polyak-Łojaciewicz inequality)

Convergence of gradient flow

$$\dot{X}_t = -\nabla f(X_t)$$

$\dot{Y}_t = -\nabla f(Y_t)$ if $Y_t = x^*$, $\nabla f(Y_t) = 0$

stationary solution

Theorem

1. *If f is α-strongly convex, then*

$$\|X_t - Y_t\|^2 \leq e^{-2\alpha t}\|X_0 - Y_0\|^2$$

in particular, if $Y_0 = x^*$: $\|X_t - x^*\|^2 \leq e^{-2\alpha t}\|X_0 - x^*\|^2$

(then $Y_t = x^*$)

2. *If f is α-gradient dominated, then*

$$f(X_t) - f(x^*) \leq e^{-2\alpha t}(f(X_0) - f(x^*))$$

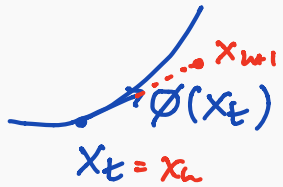
Discrete time

Discretization

Flow of a vector field $\phi\colon \mathbb{R}^n \to \mathbb{R}^n$:

$t \geq 0$

$$\dot{X}_t = \phi(X_t)$$

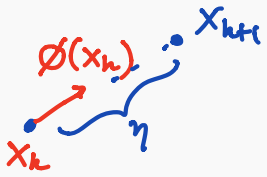


In discrete time with step size $\eta > 0$:

$k = 0, 1, 2, \ldots$

- **Forward method**

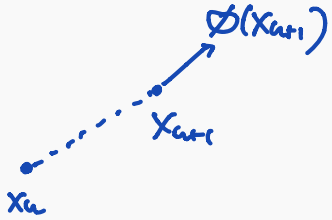
$$x_{k+1} = x_k + \eta\,\phi(x_k)$$



$$\Leftrightarrow \quad \frac{x_{k+1} - x_k}{\eta} = \phi(x_k)$$

- **Backward method**

$$x_{k+1} = x_k + \eta\,\phi(x_{k+1})$$



- Hairer, Lubich, & Wanner, *Geometric numerical integration: Structure-preserving algorithms for ordinary differential equations*, Springer, 2006

Discretization of Gradient Flow

Gradient flow:

$$\dot{X}_t = -\nabla f(X_t)$$

In discrete time with step size $\eta > 0$:

- **Gradient descent** (forward method)

$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

- **Proximal point method** (backward method)

$$x_{k+1} = x_k - \eta \nabla f(x_{k+1})$$

Example: Quadratic in 1 dimension

In $\mathbb{R}$, gradient flow of $f(x) = \frac{\alpha}{2}x^2$ is $(\alpha > 0)$

$$\dot{X}_t = -\alpha X_t \quad \Rightarrow \quad X_t = e^{-\alpha t} X_0 \longrightarrow 0 \text{ as } t \to \infty$$

- **Gradient descent**

$$x_{k+1} = x_k - \eta\alpha\, x_k = (1-\eta\alpha)\, x_k$$

$$\Rightarrow x_k = (1-\eta\alpha)^k x_0$$

$$* \quad x_k \longrightarrow 0 \quad \text{iff} \quad |1-\eta\alpha| < 1 \iff 0 < \eta < \frac{2}{\alpha}$$

- **Proximal point method**

$$x_{k+1} = x_k - \eta\alpha\, x_{k+1}$$

$$\Leftrightarrow x_{k+1} = \frac{x_k}{1+\eta\alpha}$$

$$\Leftrightarrow x_k = \frac{x_0}{(1+\eta\alpha)^k} \longrightarrow 0 \quad \text{for all } \eta > 0$$

Example: Quadratic in 1 dimension

In $\mathbb{R}$, gradient flow of $f(x) = \frac{\alpha}{2}x^2$ is

$$\dot{X}_t = -\alpha X_t \quad \Rightarrow \quad X_t = e^{-\alpha t} X_0$$

- **Gradient descent**

$$x_{k+1} = x_k - \eta\alpha\, x_k \quad \Rightarrow \quad x_{k+1} = (1 - \eta\alpha) x_k$$

$\star$ $x_k \to 0$ if $|1 - \eta\alpha| < 1 \;\Leftrightarrow\; 0 < \eta < \frac{2}{\alpha}$

- **Proximal point method**

$$x_{k+1} = x_k - \eta\alpha\, x_{k+1} \quad \Rightarrow \quad x_{k+1} = \frac{x_k}{1 + \eta\alpha}$$

$\star$ $x_k \to 0$ for all $\eta > 0$

Example: Quadratic in n dimensions

In $\mathbb{R}^n$, gradient flow of $f(x) = \frac{1}{2}x^\top Ax$ is $(A \succ 0)$

$$\dot{X}_t = -AX_t \quad \Rightarrow \quad X_t = e^{-At}X_0 \to 0 \text{ as } t \to \infty$$

- **Gradient descent**

$$x_{k+1} = x_k - \eta A x_k \quad \Rightarrow \quad x_{k+1} = (I - \eta A)x_k$$

$\star$ $x_k \to 0$ if $-I \prec I - \eta A \prec I \Leftrightarrow 0 < \eta < \frac{2}{\lambda_{\max}(A)}$

- **Proximal point method**

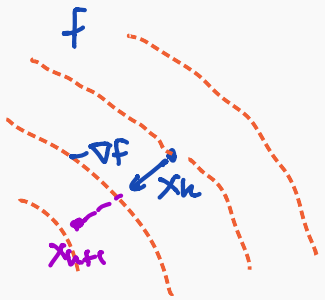
$$x_{k+1} = x_k - \eta A x_{k+1} \quad \Rightarrow \quad x_{k+1} = (I + \eta A)^{-1}x_k$$

$0 \prec (I + \eta A)^{-1} \prec I$

$\star$ $x_k \to 0$ for all $\eta > 0$

Gradient descent

Gradient descent



$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

$$= \arg\min_{x \in \mathbb{R}^n} \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\eta} \|x - x_k\|^2 \right\}$$

first-order approximation of $f(x)$ starting from x_k

- First-order method, explicit
- Greedy, descent method for small η (but not for large η)

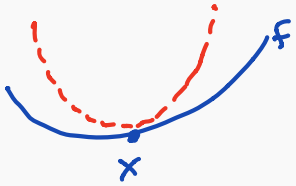
Gradient flow $\dot{x}_t = -\nabla f(x_t)$

$$-\nabla f(x_t) = \arg\min_{v \in \mathbb{R}^n} \left\{ \langle \nabla f(x_t), v \rangle + \frac{1}{2} \|v\|^2 \right.$$

Smoothness

$f : \mathbb{R}^n \to \mathbb{R}$

Recall f is L-smooth if



$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$$

$$\Leftrightarrow \quad (\nabla f(y) - \nabla f(x))^\top (y - x) \le L\|y - x\|^2$$

$$\Leftrightarrow \quad \nabla^2 f(x) \preceq LI$$

If f is α-strongly convex and L-smooth, then condition number is

$$\kappa = \frac{L}{\alpha}$$

Descent property

$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

Lemma: If f is L-smooth and $\eta < \frac{2}{L}$, then along gradient descent

$$f(x_{k+1}) \leq f(x_k) - \eta \left(1 - \frac{\eta L}{2}\right) \|\nabla f(x_k)\|^2 \quad < f(x_k)$$

Descent property

$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

Lemma: If f is L-smooth and $\eta < \frac{2}{L}$, then along gradient descent

$$f(x_{k+1}) \leq f(x_k) - \eta \left(1 - \frac{\eta L}{2}\right) \|\nabla f(x_k)\|^2$$

- In particular, if $\eta = \frac{1}{L}$, then

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2$$

- *Note:* Does *not* need convexity

Proof: By L-smoothness,

$$
\begin{aligned}
f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), \underbrace{x_{k+1} - x_k}_{-\eta \nabla f(x_k)} \rangle + \frac{L}{2} \underbrace{\|x_{k+1} - x_k\|^2}_{-\eta \nabla f(x_k)} \\
&= f(x_k) - \eta \|\nabla f(x_k)\|^2 + \frac{L}{2} \eta^2 \|\nabla f(x_k)\|^2 \\
&= f(x_k) - \eta \left(1 - \frac{\eta L}{2}\right) \|\nabla f(x_k)\|^2
\end{aligned}
$$

□

Proximal method

Proximal method

$$x_{k+1} = x_k - \eta \nabla f(x_{k+1})$$
$$= \arg\min_{x \in \mathbb{R}^n} \left\{ \underbrace{f(x) + \frac{1}{2\eta}\|x - x_k\|^2}_{:= f_\eta} \right\}$$

- Implicit method, need to solve optimization in each iteration (strongly convex for small η) if $\nabla^2 f(x) \geq -R \cdot I$, $R > 0$ then $\nabla^2 f_\eta = \nabla^2 f + \frac{1}{\eta} I > 0$ $\Leftrightarrow$ $\eta < \frac{1}{R}$

- *Always* a descent method for all $\eta > 0$

Descent property

$$x_{k+1} = x_k - \eta \nabla f(x_{k+1})$$

Lemma: For any $\eta > 0$, along proximal method:

$$f(x_{k+1}) \leq f(x_k) - \frac{\eta}{2} \|\nabla f(x_{k+1})\|^2$$

- *Note:* Does *not* need convexity, smoothness

Proof: Since x_{k+1} minimizes $f(x) + \frac{1}{2\eta}\|x - x_k\|^2$,

$$f(x_{k+1}) + \frac{1}{2\eta}\|x_{k+1} - x_k\|^2 \leq f(x_k)$$

Therefore, since $x_{k+1} - x_k = -\eta \nabla f(x_{k+1})$,

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \frac{1}{2\eta}\|x_{k+1} - x_k\|^2 \\ &= f(x_k) - \frac{\eta}{2}\|\nabla f(x_{k+1})\|^2 \end{aligned}$$

□

Convergence rate of Gradient Descent

Convergence rate under strong convexity

$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

$y_{k+1} = y_k - \eta \nabla f(y_k)$

Theorem

Assume f is α-strongly convex and L-smooth, $\kappa = \frac{L}{\alpha}$. If $0 < \eta \leq \frac{2}{\alpha + L}$, then gradient descent has exponential contraction

$$\|x_k - y_k\|^2 \leq \left(1 - \eta \frac{2\alpha L}{\alpha + L}\right)^k \|x_0 - y_0\|^2$$

In particular, with $\eta = \frac{2}{\alpha + L}$ and $y_k = x^*$

$1 - r \leq e^{-r}$

$(1-r)^{2k} \leq e^{-2rk}$

$$\|x_k - x^*\|^2 \leq \left(1 - \frac{2}{1+\kappa}\right)^{2k} \|x_0 - x^*\|^2 \leq e^{-\frac{4k}{1+\kappa}} \|x_0 - x^*\|^2$$

to get $\|x_k - x^*\| \leq \varepsilon$, need $k \geq \frac{(1+\kappa)}{2} \log \frac{\|x_0 - x^*\|}{\varepsilon} \approx \tilde{O}(\kappa)$

Essential lemma

Lemma
If f is α-strongly convex and L-smooth, then for any $x, y \in \mathbb{R}^n$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\alpha L}{\alpha + L} \|x - y\|^2 + \frac{1}{\alpha + L} \|\nabla f(x) - \nabla f(y)\|^2$$

- [Nesterov, *Lectures on Convex Optimization*, Springer, 2004], Theorem 2.1.12

Proof of Theorem

Proof: Expanding the square and using the lemma

$$\begin{aligned}
&\|x_{k+1} - y_{k+1}\|^2 \\
&\quad = \|x_k - y_k\|^2 - 2\eta \langle \nabla f(x_k) - \nabla f(y_k), x_k - y_k \rangle + \eta^2 \|\nabla f(x_k) - \nabla f(y_k)\|^2 \\
&\quad \leq \left(1 - 2\eta \frac{\alpha L}{\alpha + L}\right) \|x_k - y_k\|^2 + \left(\eta^2 - \frac{2\eta}{\alpha + L}\right) \|\nabla f(x_k) - \nabla f(y_k)\|^2
\end{aligned}$$

< 0 if η is small

If $0 < \eta \leq \frac{2}{\alpha + L}$, then

$$\|x_{k+1} - y_{k+1}\|^2 \leq \left(1 - 2\eta \frac{\alpha L}{\alpha + L}\right) \|x_k - y_k\|^2$$

Convergence rate under gradient domination

$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

Theorem

Assume f is α-gradient dominated and L-smooth, $\kappa = \frac{L}{\alpha}$.

If $0 < \eta \leq \frac{2}{L}$, then along gradient descent,

$$f(x_k) - \min f \leq \left(1 - 2\eta\alpha\left(1 - \frac{\eta L}{2}\right)\right)^k (f(x_0) - \min f)$$

In particular, with $\eta = \frac{1}{L}$,

$$f(x_k) - \min f \leq \left(1 - \frac{1}{\kappa}\right)^k (f(x_0) - \min f) \leq e^{-k\kappa} (f(x_0) - \min f)$$

$\Rightarrow$ to get $f(x_n) - \min f \leq \varepsilon$, need $k \gtrsim \kappa \log \frac{f(x_0) - \min f}{\varepsilon} = \tilde{O}(\kappa)$

Proof of Theorem

Proof: By the descent property of gradient descent

$$f(x_{k+1}) \leq f(x_k) - \eta \left(1 - \frac{\eta L}{2}\right) \|\nabla f(x_k)\|^2$$

$$\geq 2\alpha (f(x_k) - \min f)$$

By the gradient domination property

$$\begin{aligned} f(x_{k+1}) - \min f &\leq f(x_k) - \min f - 2\eta\alpha \left(1 - \frac{\eta L}{2}\right) (f(x_k) - \min f) \\ &= \left(1 - 2\eta\alpha \left(1 - \frac{\eta L}{2}\right)\right) (f(x_k) - \min f) \end{aligned}$$

□

Convergence rate of Proximal Method

Convergence rate under gradient domination

$$\begin{aligned} x_{k+1} &= x_k - \eta \nabla f(x_{k+1}) \\ &= \arg\min_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{1}{2\eta} \|x - x_k\|^2 \right\} \end{aligned}$$

Theorem
Assume f is α-gradient dominated. For any $\eta > 0$, along proximal method

$$f(x_k) - \min f \leq \frac{1}{(1+\eta\alpha)^k} \left(f(x_0) - \min f \right)$$

Proof of Theorem

Proof: By the descent property of proximal method

$$f(x_{k+1}) \le f(x_k) - \frac{\eta}{2}\|\nabla f(x_{k+1})\|^2$$

$$\geq 2\alpha(f(x_{k+1}) - \min f)$$

By the gradient domination property

$$f(x_{k+1}) - \min f \le f(x_k) - \min f - \eta\alpha\,(f(x_{k+1}) - \min f)$$

Therefore

$$f(x_{k+1}) - \min f \le \frac{1}{1+\eta\alpha}\,(f(x_k) - \min f)$$

□

Recap

For optimization $\min_{x \in \mathcal{X}} f(x)$ on $\mathcal{X} = \mathbb{R}^n$

we have studied

dynamics: gradient flow

algorithms: gradient descent, proximal method

with oracle access: gradient information

under assumption: strong convexity ($+$ smoothness)

get exponential convergence rate with $O(\kappa)$ dependence

Variations

Variations

- Optimal algorithm
- Weakly convex optimization
- Higher-order optimization
- Zero-order optimization
- ...

Optimal algorithm

Lower bound

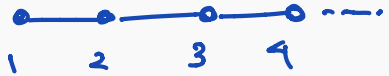
For minimizing $f : \mathbb{R}^n \to \mathbb{R}$ which is α-strongly convex, L-smooth

- If can only access $\nabla f(x)$, and output is in linear span of previous points and gradients, then lower bound is $\Omega(\sqrt{\kappa})$
- Worst-case function is quadratic (take $\mathcal{X} = \ell_2 = \mathbb{R}^\infty$)

$$f(x) = \frac{\alpha(\kappa - 1)}{8}(x^\top A x - 2x^{(1)}) + \frac{\alpha}{2}\|x\|^2$$

where A is Laplacian of path graph:

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & \ddots \\ 0 & 0 & \ddots & \ddots \end{pmatrix}$$

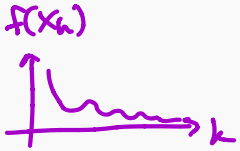


Accelerated gradient descent

$$\begin{cases} x_{k+1} = y_k - \dfrac{1}{L}\nabla f(y_k) \\ y_{k+1} = x_{k+1} + \beta(x_{k+1} - x_k), \qquad \beta = \dfrac{\sqrt{L} - \sqrt{\alpha}}{\sqrt{L} + \sqrt{\alpha}} \end{cases}$$

$$\Rightarrow \quad f(x_k) - \min f \leq 2\left(1 - \frac{1}{\sqrt{\kappa}}\right)^k (f(x_0) - \min f)$$

- Second-order method (needs two iterates)
- *Not* a descent method (can be oscillatory)



- Faster than gradient descent, achieves optimal rate!

[Nesterov, *Lectures on Convex Optimization*, Springer, 2004], Theorem 2.3.6

Convex optimization

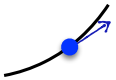
Convex optimization

$f : \mathbb{R}^n \to \mathbb{R}$ is *(weakly) convex* if $\nabla^2 f(x) \succeq 0$

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Gradient flow: $f(X_t) - \min f \leq O\left(\frac{1}{t}\right)$
- Gradient descent: $f(x_k) - \min f \leq O\left(\frac{1}{\eta k}\right)$ if f is smooth
- Optimal rate: $\Theta\left(\frac{1}{\eta k^2}\right)$, Nesterov acceleration

Gradient Flow



$$\dot{X}_t = -\nabla f(X_t)$$

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

[Su, Boyd, Candes '14]

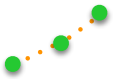
Accelerated GF

$$\ddot{X}_t + \frac{3}{t}\dot{X}_t + \nabla f(X_t) = 0$$

$$O\left(\frac{1}{t^2}\right)$$

f convex

Gradient Descent

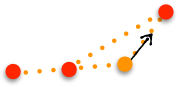


$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

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

[Nesterov '83]

Accelerated GD



$$x_{k+1} = y_k - \eta \nabla f(y_k)$$

$$y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1})$$

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

f convex,
$\nabla^2 f$ bounded

Higher-order optimization

Higher-order optimization

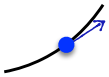
If we have *higher-order derivatives*, then we can get faster rate

- In continuous time, corresponds to rescaled gradient flow
- In discrete time, can do higher-order gradient descent or rescaled gradient descent
- Can be accelerated, variational view via Bregman Lagrangian

⋆ [Wibisono, Wilson, Jordan, *A Variational Perspective on Accelerated Methods in Optimization*, PNAS, 2016]

⋆ [Wilson, Mackey, Wibisono, *Accelerating Rescaled Gradient Descent: Fast Optimization of Smooth Functions*, NeurIPS, 2019]

Higher-order convex optimization



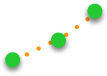
Rescaled Gradient Flow

$$O\left(\frac{1}{t^{p-1}}\right)$$

f convex

Accelerated Flow

$$O\left(\frac{1}{t^{p}}\right)$$



Higher-order Gradient Descent

Rescaled Gradient Descent

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

Acceleration

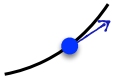
f convex, p-smooth

Accelerated Descent Method

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

Dynamics for optimization

$$\min_{x \in \mathbb{R}^n} f(x)$$



Rescaled Gradient Flow

$$\dot{X}_t = -\frac{\nabla f(X_t)}{\|\nabla f(X_t)\|^{\frac{p-2}{p-1}}}$$

- First-order dynamics
- Greedy descent flow
- Related via space regularization

$$O\left(\frac{1}{t^{p-1}}\right)$$

Accelerated Flow

$$\frac{d}{dt}\nabla h\left(X_t + \frac{t}{p}\dot{X}_t\right) = -t^{p-1}\nabla f(X_t)$$

- Second-order dynamics
- Comes from Bregman Lagrangian
- Related via speeding up time

$$O\left(\frac{1}{t^{p}}\right)$$

f convex

Zero-order optimization

Zero-order optimization

What if we only have *zero-order* information (function value)?

- Estimate gradient via function values at random points:

$$\nabla f(x) \approx \frac{f(x + \eta Z) - f(x)}{\eta} Z$$

  where $\mathbb{E}[Z] = 0$, $\mathrm{Cov}(Z) = I$

⋆ Duchi, Jordan, Wainwright, & Wibisono, *Optimal Rates for Zero-Order Convex Optimization: The Power of Two Function Evaluations*, IEEE Transactions on Information Theory, 2015

Other variations I

- Non-Euclidean geometry
  - Mirror descent using Bregman divergence (Hessian manifold)
  - Multiplicative weight update as mirror descent with entropy regularizer (Fisher metric)
- Affine-invariant optimization
  - Newton's method, interior point method
  - Superfast convergence under *self-concordance*, barrier property
- Non-smooth optimization: f not differentiable
  - Subgradient method
  - Smoothing

Other variations II

- Stochastic optimization: $f(x) = \mathbb{E}[F(x;\theta)]$
  - Stochastic gradient descent
  - Coordinate descent
  - Variance reduction
- Composite optimization: $\min_{x\in\mathbb{R}^n} f(x) + g(x)$
  - Forward-backward
  - ISTA, FISTA
- Approximations of proximal method
  - Optimistic, extra-gradient methods
  - Important for min-max games and variational inequalities
- ...

References

References

- Hairer, Lubich & Wanner, *Geometric numerical integration: Structure-preserving algorithms for ordinary differential equations*, Springer, 2006
- Karimi, Nutini & Schmidt, *Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition*, ECML, 2016
- Beck & Teboulle, *A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems*, SIAM Journal of Imaging Sciences, 2009

References

- Boyd and Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004, available at `https://web.stanford.edu/~boyd/cvxbook`
- Nesterov, *Lectures on Convex Optimization*, Springer, 2004
- Nocedal and Wright, *Numerical Optimization*, Springer, 2006
- Vishnoi, *Algorithms for Convex Optimization*, available at `https://convex-optimization.github.io`

  CPSC 463: *Algorithms for Continuous Optimization*, Fall 2021
- Recht, *Optimization*, Big Data Bootcamp at Simons Institute, 2013, `https://simons.berkeley.edu/talks/ben-recht-2013-09-04`
- Boyd, *Convex Optimization*, Lecture videos for EE 364A at Stanford, 2008, `https://www.youtube.com/watch?v=McLq1hEq3UY`